DOCUMENT RESUME

ED 106 327 32 TM 004 442

AUTHOR	Durost, Walter N.
TITLE	A Description and Evaluation of the Statewide Testing Program ir New Hampshire in 1968-69 and 1969-70 Under the Sponsorship of Title I and the Significance of the Data Obtained For Evaluation with this Activity.
INSTITUTION	New Hampshire State Dept. of Education, Concord.
SPONS AGENCY	Bureau of Elementary and Secondary Education (DHEW/OE), Washington, D.C. Div. of Compensatory Education.
PUB DATE	Mar 71
NOTE	84p.; Document not available in hard copy due to marginal legibility of original document; Product of the Test Service and Advisement Center, Lee, N.H.
EDRS PRICE	MF-$0.76 HC Not Available from EDRS. PLUS POSTAGE
DESCRIPTORS	Academic Achievement; Achievement Tests; Comparative Testing; *Compensatory Education Programs; Group Intelligence Tests; Intelligence; *Program Evaluation; Program Improvement; School Districts; *State Programs; *Testing Programs; Test Interpretation; *Test Results
IDENTIFIERS	Elementary Secondary Education Act Title I; ESEA Title I; *New Hampshire; New Hampshire Statewide Testing Program; State Testing Programs

ABSTRACT

The New Hampshire statewide testing program was implemented to provide a data base for the evaluation of the effectiveness of Title I projects as required by Federal law. To accomplish this objective, achievement and intelligence tests were administered to children in Title I projects and regular programs in four elementary grades--2, 4, 6 and 8. Thus the performance of children in both programs could be analyzed and compared. The information collected during the 1968-69 program was used as a basis for modifying and improving the 1969-70 program. Test results, statewide analysis and interpretation of the data are presented. (EVH)

A DESCRIPTION AND EVALUATION OF THE STATEWIDE TESTING PROGRAM IN NEW HAMPSHIRE IN 1968-69 AND 1969-70 UNDER THE SPONSORSHIP OF TITLE I AND THE SIGNIFICANCE OF THE DATA OBTAINED FOR EVALUATION WITH THIS ACTIVITY

PREPARED BY: WALTER N. DUROST, Ph.D., DIRECTOR

TEST SERVICE AND ADVISEMENT CENTER

Completed under contract of March 13, 1971.

New Hampshire Statewide Testing Program Evaluation
1968-69 and 1969-70

List of Contents

N. H. Statewide Testing Program Evaluation - List of Contents (Continued)

A Description and Evaluation
of the STATEWIDE TESTING PROGRAM in NEW HAMPSHIRE
in 1968-69 and 1969-70
Under the Sponsorship of Title I
And the Significance of the Data Obtained
For Evaluation With This Activity

SECTION I

Introduction

Research is a dirty word to many and an ambiguous word to those who endeavor to carry out activities so named. Was Edison doing research when he was experimenting with filament materials for the electric light bulb? The writer feels that he was. His efforts represented a planned attack on the problem with more or less precise specifications of the capabilities of the material being sought. His efforts, however, could never have been subject to PERT analysis and he never would have gotten a government contract.

In many ways this study faced the same dilemma. All involved had a fairly clear picture of what needed to be done; namely, to make use of objective testing and data collection procedures plus appropriate analysis techniques to arrive at a value judgment as to the "goodness" of the Title I effort in this state. This objective was hedged about by many restrictions of regulation, administration and philosophy, some of which were almost directly incompatible with the goal stated above. Much was done toward reaching the goal, much more could have been done under other circumstances. The following pages constitute a real effort to contend with all the difficulties and present eventually some kind of meritorious report. A good place to start the report is with a statement of the purposes and procedures in setting the Title I evaluation program in motion in New Hampshire.

Provisions of the Federal Law as Regards the Selection of Students for Title I Projects

Perhaps some information at this point concerning the development and implementation of Title I as a major part of ESEA will provide a background which will clarify some of our problems in regard to the subsequent data analysis.

Money for Title I programs is allocated to each state in terms of the number of families, county by county, falling below the national poverty level plus some other considerations, such as number of families receiving Aid for Dependent Children and number of children in foster homes. All these data are used, along with similar information from the other states, to determine the proportion of available funds to come into this state. However, the Title I office within the State Department of Education determines the distribution of money by school district as against county, again being guided by the economic considerations as listed above. In other words, the State Department's allocation of funds to a district is strictly in accordance with the statistics concerning the number of families qualifying in that district as defined above.

School districts in New Hampshire, like all other states in the country, are required by Title I regulations to designate target schools except under certain conditions. Basically, this provision is inoperative in New Hampshire in a majority of the school districts because there exists only a single attendance area. In the larger cities, such as Dover, Portsmouth, and the like where there are multiple attendance areas, certain schools are designated as target schools. In such instances, Title I projects are confined solely to these schools.

Subsequently, each school district is responsible for submitting projects to the State Department Title I office for approval as a basis for the specific allocation of monies to fund these projects.

Each school district proposal is expected to state very explicitly the grades involved, expense for personnel and material, and, finally, the method to be used to evaluate outcomes. Such evaluation is mandatory according to Federal regulations.

These Federal regulations favor objective testing as the primary basis for evaluation and make some further stipulations in regard to progress as measured in grade equivalents, which make no sense from a measurement point of view.

Obviously, when a very large proportion of the students in Title I at all grades are involved in reading projects, it makes sense to use a reading test as part of an evaluation and possibly even as the basic instrument for assessing outcomes. This heavy emphasis on reading, incidentally, is a rather general characteristic of Title I projects throughout the country.

The U. S. Office of Education reflects this emphasis by its recently activated project for equating the major reading tests avilable through commercial channels and restandardizing one of them (Metropolitan) as an anchor test.

School districts have the option of selecting the tests to be used within the district for evaluation of their own Title I programs. Some choose to use tests consistent with those used throughout the school district for all pupils, even if these are not particularly suitable for the purpose. Consquently, the number of Title I children tested Fall and Spring in New Hampshire with the state designated tests is substantially lower than the total number of children in Title I in the respective grades. This must be considered an exercise

of democracy at the expense of the common good.

As a matter of fact, the proportion of Title I students in the tested Spring sample is only slightly more than 50% of known Title I cases, i.e., cases for whom both "IN" and "OUT" cards are available. For example, in Grade 4, about 100 pupils were enrolled in Title I projects in large districts which chose NOT to test in the statewide testing program.

In addition to the attrition due to the failure of a district to take part in the state program at all, account must be taken of the difficulties involved in matching Fall and Spring tested cases, as described elsewhere, as well as other causes of attrition which are hard to assess as to total effect. This includes such factors as absence from school for one test or another or for one segment of the battery administered in the Fall or in the Spring. Tested cases used in this analysis included only those for whom we had test information on the five skills tests used; namely, Word Meaning Paragraph Meaning, Arithmetic Computation, Arithmetic Concepts and Arithmetic Applications.

Another serious source of error is in the failure of some school districts to test in the Spring even though they did participate in the Fall testing program.

The analysis in Section IV is based upon completely tested cases only, matched for Fall and Spring testing. Only in Table IV-1 of this section, concerned with determining the percent of children in each category of Title I projects, does the report utilize the complete set of "IN" cards.

From all of the above, it must be amply clear that it would be very difficult to defend the Title I completely tested sample as being representative of all Title I cases in the State at any grade level. It would be difficult even to defend the sample statistically as being representative of those pupils in academically oriented programs or, even more specifically, in corrective and remedial programs in reading.

On logical and experiential grounds, it SEEMS representative. Even if it is not precisely so, it still constitutes a viable population of Title I pupils, defined by all the descriptive information reported here. It is therefore quite legitimate to report for this defined population the results of the analysis of the "IN" and "OUT" card data. At least the two aspects of the analysis are comparable. 1/

One important fact concerning the use of Title I funds is often overlooked. These monies are not intended to defray expenses ordinarily provided for in the regular school budget. For the sake of illustration, let us assume that we are discussing an urban community with designated target schools. One or more such target schools may be receiving money for a remedial reading program but OTHER NON-TARGET SCHOOLS may be just as needful of these services. Still the district cannot use Title I funds to finance a similar remedial reading program in these other schools, even though badly needed, because such schools do not satisfy the criteria for Title I aid established in the Law and enabling regulations.

Considerations such as this highlight the importance of the "IN" and "OUT" cards as basic control documents. Test results relating to Title I pupils in this study are reported only for documented Title I cases. It is unfortunate that up to 50% of these Title I cases, so identified, were not tested for reasons specified above. More might have been tested if some coercion had been used. The Title I office did not wish to be put in the position of dictating the method of evaluation (i.e., by means of a particular standardized test) to be used for all projects, especially if the project, such as a speech therapy facility, was clearly not subject to evaluation by a standardized achievement test.

No coercion was employed to obtain participation if a good district-wide testing program would have been impaired by insisting that the Stanford Achievement Test be given to ALL Title I students. There is indeed a serious dilemma here! How can a truly representative evaluation for Title I enrollees be produced if no central office has the authority to prescribe the conditions and instruments to be used?

The other "horn" of the dilemma relates to the reasonableness of any statewide evaluation of the Title I program by means of standardized tests that does not also involve a district by district evaluation, taking due account of the degree of appropriateness of the instrument in terms of the local project design. In such total population studies, everything tends to be reduced to the level of mediocrity, whereas the truth is that some districts excel and some fail miserably, even in a limited area such as reading.

Let it be abundantly clear that this writer has no recourse but to report results on the samples tested Fall and Spring and only touch marginally on the relative goodness of the effort from district to district. This is a source of no little frustration. District averages for Title I cases are not a satisfactory answer. Aside from the misleading nature of such averages, the number of Title I cases is so small in many districts that one would have to resort to weighting of some sort to make comparisons fair.

Finally, averages are of little practical value for comparison purposes unless the criteria for selection of cases in terms of educational need are

1/ These are the questionnaire aspect versus the statistical aspect.

highly standardized and the type of treatment is at least roughly specified.

The 1968-69 Testing Program in New Hampshire

Early in 1968 a group of people in the State Department of Education, headed by Mr. James Carr, State Director of Guidance, working closely with the Title I office, Mr. William Sterling, Director, planned a statewide testing program, one of the main purposes of which was to provide a data base from which to evaluate the outcomes of Title I projects in accordance with Federal law. While the author of this report was consulted sporadically, there was no official relationship at the beginning of the program, especially in the selection of the tests to be used. Subsequently, Test Service and Advisement Center was asked to review, summarize, expand and interpret the test data for the '68-'69 program and subsequently to do the same for a repeat program in '69-'70. This report is largely concerned with this second evaluation and will include the presentation of certain data which has not previously been made public.

The grades involved in both the '68-'69 and '69-'70 programs were 2, 4, 6 and 8. Alternate grades were chosen because it was not economically feasible to test all grades, although Title I programs were going on in all grades. The basic plan was to test as comprehensively as possible in these four grades and then to evaluate the results of the Title I programs in the state in terms of a repeat testing program in the Spring, using the same test, even the same forms. The plan to use the same forms in the Spring was not considered to be advisable by this writer, although in retrospect it seems not to have made much difference, as will be seen by making some comparisons of the '68-'69 and '69-'70 data.

The Plan for Testing

Although this is not the place to go into detail as to the plan adopted for implementing these programs, it is essential to point out that, for the two years involved, the State contracted with Harcourt Brace Jovanovich (then Harcourt, Brace & World, Inc.) to conduct the scoring and the analysis of the data through its Programs and Services Division. This arrangement was arrived at only after Mr. Carr and others involved at the time consulted, personally, with both Harcourt and Educational Testing Service as to the best program possibilities. The original planning contemplated a three-year cycle.

The tests chosen were as shown in the table below:

	Otis-Lennon Mental Ability Test	Stanford Achievement Test
Grade 2	Elem. I, Form J	Prim. I, Form X
Grade 4	Elem. II, Form J	Inter. I, Form X
Grade 6	Elem. II, Form K	Inter.II, Form X
Grade 8	Inter., Form J	Advanced, Form X

All materials in this program were scorable by means of the optical scanning equipment available at the Measurement Research Center at Iowa City, but the batteries used at Grade 2 consisted of scoreable booklets which were consumed in the process. The booklets for the remaining grades, 4, 6 and 8, were reusable and were left in the schools to be used over again, as needed.

In the Spring, the Stanford Achievement Battery was to be repeated. In every case, the same form was used except for Grade 2 in which Form W of the Primary II Battery of Stanford was substituted for Form X. Note: This was a change of both form and battery. The pattern of using an alternative form for Spring testing in all grades was conceded to be the appropriate pattern but considerations of economics prevailed and the same form was used over again in Grades 4, 6 and 8.

This procedure, as outlined above, was repeated, so far as the tests are concerned, in the 1969-70 program.

Evaluation of the 1968-69 Program

In evaluating the 1968-69 program, it must be remembered that a program of this magnitude had never before been undertaken in New Hampshire and although the Division of Programs and Services at Harcourt, under the contract, was responsible for providing technical and professioanl leadership, the amount of this leadership was minimal. The responsibility for this minimal involvement, however, is a shared responsibility since there was no great insistence on the part of the State Department personnel that such professional assistance be provided beyond the barest outline of procedures.

The evaluations which follow must be recognized to be those of the writer, who was called in to provide technical and professional assistance in a further analysis of the data beyond the point carried out by MRC. The writer's task was to try to critique the 1968-69 program to suggest ways in which it might be improved in 1969-70. Every effort has been made to be objective and fair to all persons concerned, but the net judgment must be that the '68-'69 program did not yield all of the data needed to carry out the purposes of the program as stated and the actual administration of the program left much to be desired in terms of quality of test administration, document preparation for scoring and adherence to schedule.

All tests were to be administered throughout the state during a more or less uniform period in mid-October, but actually the test administration was carried on over a longer period of time than was planned. There was no central office supervision to insure that the tests were properly administered, although there is little evidence that they were grossly misadministered at the local level.

The return of the data to MRC for scoring, which should have been highly uniform, both in terms of time in order to facilitate processing, and in terms of preparation of essential coding information, was very poor. Many schools were very late in returning their booklets to MRC. Furthermore, preparation of header sheets was sloppy, personal ID information was often missing and the protocols were not properly arranged to facilitate rapid processing and return of reports. Consequently, the reports did not come in for many weeks after the tests were administered and actually some community reports did not get back to the schools until months after the test administration.

The analysis at the Measurement Research Center, Iowa City, as specified to them by Harcourt, was fairly comprehensive. This consisted of the basic MRC service, plus options which could be chosen by the local communities as they wished to do. Thus, every community received class rosters by school and class, giving the results of both the Otis-Lennon tests and the Stanford Batteries involved. For Otis-Lennon, scores and IQs were reported together with local stanines. For Stanford, grade scores, grade equivalents, percentile ranks and stanines were reported. In addition, for the Stanford test, item analysis data were made available for the entire state and separately by community where the communities wished to have this information. These statwide item analysis data were handed over to the members of the Division most directly concerned. For example, the item analysis data for the state as a whole in the field of mathematics went to the consultant in mathematics, the information on science went to the consultant in science, etc. So far as the writer knows no consistent, systematic use was made of this information at the State level. It is impossible to know, of course, to what extent communities made use of the data.

Workshops were held around the state on a regional basis but these were one-day workshops. The morning session was for administrators and the afternoon session for teachers. The coverage of the test information was superficial and so far as the writer could observe afterwards, by making some study of this aspect of the program, the impact was minimal. On the other hand, it must also be said that there was no substantial dissatisfaction indicated, perhaps because there was little sense of need for such in-service training.

Identification of Title I Cases

The identification of Title I cases was done by means of the allocation of one space on the "Other Data" section of the MRC answer sheet or ID page of the Primary I Battery to indicate that the pupil was in a Title I project. Unfortunately, the tests were given too early in the year for this determination to have been made in every case. Many pupils, thought to be qualified for Title I help, were so indicated, even though subsequently they did not take part or only stayed in the program briefly, while many pupils who did subsequently take part in the program were not identified. Many more were originally identified as Title I than actually were served by the schools.

Furthermore, there was no differentiation as to the category of Title I projects involved; the child was merely designated as a Title I subject. This failure to identify type of projects was quickly picked up at the first group meeting, during the summer after testing, called to discuss the results of the statewide testing and this constituted one of the major directions for change in the 1969-70 program.

Outcomes for 1968-69

The '68-'69 data returned to the communities by MRC provided the means for the local community to make use of the data not only in connection with Title I but with all children who took the test. Since the sample of schools and school districts participating was good, this meant a very substantial majority of children enrolled in these particular grades in the state were tested. No systematic attempt was made to follow up the extent to which the data were used and subsequent evidence makes it very clear that lack of knowledgeable supervision in the specific area of test utilization at the local level diminished greatly the effectiveness of the testing program. School districts in New Hampshire do not routinely have a person or persons on their staff designated as expert in the field of testing, application and utilization of test results for the improvement of instruction.

The Involvement of Test Service and Advisement Center

The writer, as the proprietor and director of Test Service and Advisement Center, was asked to coordinate the further analysis of the data over and above that done by MRC, working in cooperation with the Bureau of Educational Research and Testing Services of the University of New Hampshire. The intent was to maximize the utilization of the data to the extent possible on an "after the fact" basis. The effort to do this was a very frustrating experience in which the intent and purpose of the assignment was nearly nullified by the laxity with which the tests were administered.

In the first place, improper or incomplete ID information made it impossible to compare and collate the information from the mental ability test and the achievement test. Contrary to instructions, in many cases, full information concerning sex and chronological age was not provided for every student, nor was the student's name properly coded on all answer sheets. A flagrant violation of this coding process was the use of nicknames in place of the full legal name of the child. Thus, the early decision to develop an intrinsic code based upon selected letters in the child's name,

plus birthdate, plus sex with due account of whether or not the child was a twin was completely frustrated and a very substantial proportion of the cases were lost.

The cooperation of the Bureau of Educational Research and Testing Services at the University also left much to be desired, partly because of failure to pre-plan and thus to evaluate the amount of work that was involved in doing this analysis for so many grades. Programming was a problem, since extant programs were not sufficient to handle all of the analysis specified by the Test Service and Advisement Center, complicated by major changes in the equipment available at the Computation Center. In spite of all of these difficulties, including very poor service from the Measurement Research Center in returning the data tape in proper condition and within a reasonable time, some significant information was derived and provided to local officials for action.

The Spring Program

In the Spring of the year (for the most part of May) the Stanford Achievement Tests were readministered at the local level to children who presumably were in Title I. Local school districts were allowed to order as many tests as they wished and no attempt was made to establish the fact the children so tested were indeed in Title I. The amount of test materials purchased under this arrangement greatly exceeded the number of children enrolled in Title I and, consequently, the results which came back from MRC were virtually worthless as a basis for evaluating Title I outcomes.

Summary and Conclusions for 1968-69

From the above, it would appear that the program for '68-'69 was not too successful. Yet it is a recognized research principle that a negative result is oftentimes as significant as a positive one for indicating need for change. From this point of view, the '68-'69 program was successful. Furthermore, it did provide some very essential baseline data in both the areas of intelligence and achievement in terms of which some significant evaluation of educational output could be made for the State even though not for Title I. Out of this experience came planning for the '69-'70 program which promised radical changes in this situation.

1. Basically, the statewide data reported was adequate insofar as it went but provided for no systematic comparison of capacity and achievement to see to what extent children were working up to their optimum level, assuming for the moment that the mental ability test was adequate to determine this level.

2. The Stanford Achievement Test provided no national normative data for Spring that could be depended upon, especially in the case of Grade 2 where two different forms were used, Form X in the Fall, Form W in the Spring, as well as two levels. Grade equivalents, as the author has noted in a number of places and under different circumstances, are completely unsatisfactory for measuring the amount of learning taking place over the relatively short period of seven months. Grade equivalents for Stanford were based upon averages of grade groups tested in March of EACH school year. Thus annual, not school year, data were available for each grade in the national standardization program for determining the median scores of children tested at Grade 1, Grade 2, Grade 3, Grade 4, etc. These grade medians were plotted in terms of raw scores transformed to a single base for the grades involved and a continuous curve was drawn through the plotted points. From this, grade scores were derived which were used as a basis for tying together forms. These grade scores and the grade equivalents, elsewhere given, were identical. When they were called grade scores, the decimal point was eliminated; with the decimal point in they became grade equivalents.

Increment in score from subject to subject always varies greatly, depending upon the degree of specificity of instruction in the local school. Grade equivalents in arithmetic, for example, have an entirely different significance than grade equivalents in reading or vocabulary because the effect of the environment is very much greater in the latter two instances than in areas as specifically related to school learning as arithmetic.

Furthermore, an examination of the data not only for this State but for many other communities and administrative units around the country seemed to indicate quite clearly that the Stanford norms were not truly representative of national achievement. Instead it seemed quite clear that the population of students tested for norm purposes was somewhat above average in mental ability. This was evident from an examination of the Stanford accessories as well as from the New Hampshire data from Otis-Lennon which showed the state group to be slightly above average in ability but seriously below the national norms on several Stanford tests. This is not terribly serious if local norms are used, as was strongly advised by the writer. These were provided by MRC in terms of state stanines and local stanines and percentile ranks.

An even more serious situation existed in Grade 2 due to the fact that a different battery level was used in the Spring as compared to the Fall, thus making almost impossible a comparison of the scores to determine the amount of learning taking place since the content was different.

It occasionally happens that a test which is adequate for use in the Fall is not adequate for use in the Spring of the same school year and, in this particular case, the situation was exaggerated because the Primary I Battery of Stanford used in the Fall was itself much too easy for many of the children taking the test, resulting in negatively skewed distributions, i.e., the piling up of scores at the upper or high score end of the scale.

3. Great delays were encountered in getting the tape from MRC and the tape, when received, was not wholly satisfactory. Program difficulties and scheduling difficulties at BERTS rendered the results less and less useful as the time between testing time and the return of data to the communities from the supplemental analysis was lengthened.

It would be completely senseless to try to place the major part of the blame for the failures involved in the '68-'69 data on any organization or group. It makes much more sense to take a look at these data in terms of what they did contribute to our knowledge about New Hampshire pupils and also to our awareness of the need for change in the program for the next year to improve the output. Since this report is limited as to length, no further analysis or reporting of '68-'69 data will be undertaken. Instead the concentration will be on the program for '69-'70 which was improved in major ways, although still, in the writer's opinion, falling far short of an ideal program for the purpose intended; namely, the evaluation of Title I outcomes.

SECTION II

Steps Taken to Improve The 1969-70 Statewide Testing Program And the Results of These Efforts

After a number of planning conferences, involving members of the Department of Education as well as the Title I staff, it was decided to continue with the 1969-70 program in accordance with the original plan for a three-year cycle. Mr. James Carr, State Director of Guidance, was involved in these discussions although he subsequently was away on leave for graduate study and was not involved in the 1969-70 program. This left a void which was filled by involving the Title I staff far more intimately with the testing program than had been true before.

As a result of the initial conferences, prior to any commitment to Harcourt, Brace & World for MRC service, an agreement was reached to strengthen the program in a number of ways.

1. Nine Title I project categories plus a tenth general category were developed. This was tried first for the 1968-69 program but was not effective because of inadequate information. These categories were the result of careful study of the project applications which had been approved during the previous year by the Title I staff.

2. To implement the collection of this information in 1969-70, the writer and the staff of the Title I office developed two IBM card questionnaires; the first of which was labelled the "IN" card or registration card, and the other the "OUT" card or termination card. A photocopy of these two cards is attached. These instruments provided the medium for collecting information concerning the distribution of children according to the type of Title I project in which they were involved. It was known, for example, that a very large number of Title I students were involved in corrective reading programs of one kind or another, but these data had not been quantified, nor was there any specific information concerning the number of cases involved in programs other than reading.

Furthermore, it was considered very desirable to determine to what extent Title I children remained in the project to which they were assigned for the full year or stayed only for some period less than a school year. All of this information, plus other information as may be seen by examining the cards themselves, became available through the medium of the "IN" and "OUT" cards.

These cards were distributed by the Title I office with instructions to the local communities to complete an "IN" card for each child taken into the program at whatever time this occurred and at the termination of his involvement in the project to complete an "OUT" card.

A candid evaluation of the functioning of these cards indicates that they have contributed enormously to the body of information about Title I children. They reveal some discrepancies about the number of cases in Title I enrolled versus the number of tested cases. Fuller discussion of the results of the "IN" and "OUT" card date analysis comes at a later point in the report. It is important, however, to point out here that the completion of the "IN" and "OUT" cards has not quite reached the goal intended at the time they were devised of providing, in addition to data for statistical analysis, an immediate sight file in the Title I office of all enrollees very soon after the school year starts. Problems have been encountered in distributing and collecting the cards and it is felt that there is room for improvement in this area.

3. Since the supervisory unions did not have any trained personnel to handle the testing program, it was felt important to establish a chain of command within each supervisory union so as to delegate responsibility on a pyramidal basis, with one person at the top being clearly responsible for the functioning of the program at the local level. Therefore, each supervisory union was requested to select an individual, presumably someone that had some previous training in measurement or some inclination to work with such data, to act as the supervisory union coordinator. Within the supervisory union responsibility was delegated down the chain of command to the principals of the buildings and to the teachers for seeing to it that the tests were administered on schedule, that the protocols were properly cleaned up, and that ID information was complete at the time of shipment to MRC for scoring. This was to include such things as accur-

IN-IN

NAME ..... (Last) ..... (First)

BIRTH DATE ..... (Year) ..... (Month) ..... (Day)

SEX: BOY..... GIRL..... TWIN: YES..... NO.....

GRADE: K 1 2 3 4 5 6 7 Grades 8-12 Non Graded (Circle one)

SCHOOL ..... TOWN .....

CLASSROOM TEACHER .....

ENTERING DATE ..... (Year) ..... (Month) ..... (Day)

NATURE OF PROJECT: 1 2 3 4 5 6 7 8 9 0 (See instructions; circle one)

IS THIS STUDENT INVOLVED IN ANOTHER TITLE I PROGRAM: YES.... NO....

NUMBER OF HOURS PER WEEK STUDENT IS INVOLVED IN TITLE I PROGRAM:
1 2 3 4 5 6 7 8 9 Full Time (Circle one)

INSTRUCTOR: (X one)

A. Regular Classroom Teacher only .....(1)

B. Outside Consultant plus Teacher .....(2)

C. Outside Person or Agency .....(3)

D. Special Teacher: Language.....(4), Reading.....(5), Speech.....(6), Math.....(7), Guidance.....(8), Aide.....(9), Other.....(10)

WAS STUDENT IN A TITLE I PROGRAM LAST YEAR?
Yes.....(1) No.....(2) Don't Know.....(3)

OUT

NAME ..... (Last) ..... (First)

BIRTH DATE ..... (Year) ..... (Month) ..... (Day)

SEX: BOY..... GIRL..... TWIN: YES..... NO.....

GRADE: K 1 2 3 4 5 6 7 Grades 8-12 Non Graded (Circle one)

SCHOOL ..... TOWN .....

CLASSROOM TEACHER .....

ENTERING DATE ..... (Year) ..... (Month) ..... (Day)

TERMINATION DATE ..... (Year) ..... (Month) ..... (Day)

NATURE OF PROJECT: 1 2 3 4 5 6 7 8 9 0 (See instructions; circle one)

REASON FOR TERMINATION: (X one)

A. Satisfactory progress – need met.....(1)

B. Left School.....(2) C. End of School Year.....(3)

D. Program not meeting student's needs.....(4) E. Other.....(5)

SUCCESS OF CHILD IN THIS PROGRAM: (In proportion to potential improvement) (X one)

A. Excellent progress – program of great benefit.....(1)

B. Modest progress – could have accomplished more.....(2)

C. Minor change – program did not meet basic need.....(3)

D. No real benefit.....(4) Why not? .....

.....

ate completion of the header sheets, arrangement of the answer sheets in alphabetical order within class, the removal of unscoreable sheets from each class especially those where a great many double marks appeared, where the marks were faint and probably not scoreable, or where very large numbers of items had been omitted so that it was obvious that the test was clearly invalid.

Especially great emphasis was put on the need to have accurate ID information for each child and an attempt was made to have the child's name coded in a standard fashion in order to make use of a self-generating code for each pupil as described earlier. A preliminary experiment was carried out with data from the 10th grade statewide program administered and scored by the UNH Bureau of Educational Research and Testing Services, in which Digitek answer sheets (far better designed for the purpose) had been used. The results of this analysis showed that only a tiny fraction of the total failed to complete the ID information accurately, amounting to some 20 to 25 cases out of more than 10,000 tested. This encouraged us to believe that we might get similar results at the lower grades.

However, the very inadequate design of the Stanford answer sheets and the fact that the Stanford sheets provided no space for coding birthdate, left us completely dependent for this information on the Otis-Lennon answer sheet. These design inadequacies interacting with carelessness on the part of the teachers in supervising the coding of this information by pupils, resulted in data that proved not to be very useful in establishing such a pupil code. Thus, the matching of fall and spring data for the Title I children tested was again frustrated, to say nothing of a random sample of children across the state who were also tested. These failures complicated the analysis to the point where this constituted a major stumbling block in all subsequent studies, seriously eroding the basic validity of the data herein reported.

In an early meeting of the testing supervisors from the supervisory unions, an attempt was made to get the schools to agree to make use of Social Security numbers for coding purposes, but this proved to be totally unsuccessful, although some supervisory unions did indicate a serious interest in this possibility for the future.

The matter of identifying pupils by code number within the State of New Hampshire for testing purposes still remains an unsolved problem to which the State Department of Education must, at some time, frankly face up to if it ever is to be possible to collect data concerning individual students on a cumulative basis for follow-up purposes. This, however, is not a major consideration of this report, since in this report we are concerned almost solely with the determination of the effectiveness of the Title I program in 1969-70.

4. An attempt was made to set up specific dates within which the testing program would be completed, the protocols sent from the schools to the supervisory union offices, and finally shipped from the supervisory union offices to MRC. It was our intent to keep accurate account of the dates of the receipt of the material in the supervisory union offices and the arrival dates of this material at MRC but this did not work out in practice. Many communities were late in returning their data to MRC and some failed to return them according to the specified method, namely parcel post, special fourth class mail, special handling. Consequently, the date originally set up for processing our data at MRC was missed, making it necessary for them to work us into their schedule when they could, after receiving word from the State Department of Education to begin scoring.

Quality of the MRC Service

During the 1968-69 testing program, the quality of evaluative material returned by MRC, as specified by Programs and Services of Harcourt, Brace & World, was quite adequate, with one or two minor exceptions.

The contract originally negotiated indicated that the same service would be maintained for the same price over a period of three years. However, after conversations during the summer subsequent to the 1968-69 program, there arose a serious misunderstanding between Harcourt, Brace & World, Inc., and the State Department of Education - especially the Title I office - with the result that the information returned for the 1969-70 program was substantially lacking in the degree of completeness that characterized the 1968-69 program.

Moreover, the service from MRC certainly was not improve although who was at fault in this respect it is hard to say. To some extent, at least, it was the failure on the part of the State Department in making it clear that the same service was expected and the subsequent failure of the representatives of Harcourt, Brace & World to write specifications to insure this same degree of completeness. However, there was no consequent change in price for this less adequate service!

Because of the difficulty in getting the IBM cards shipped intact from MRC in 1969, it was decided to go to magnetic tape. However, the magnetic tape was not shipped from MRC until all other aspects of the contract had been completed. The delay, amounting to some months, in getting the tape to New Hampshire tremendously retarded the analysis of the data by the Bureau of Educational Research and Testing Services of the University of New Hampshire. There were programming problems involved also because of incomplete specifications from MRC and errors in the tape. Although the analysis requested from BERTS was essentially simple and similar programs had been carried out routinely in many places around the country before, nevertheless delays of substantial length were incurred. Furthermore, because of the experimental nature of the program, new ways of looking at the

dats kept occurring to the writer and to the members of the staff of Title I, necessitating expanding the analysis beyond the points originally contemplated. The problems of lost and poorly shipped materials continued to plague the program in spite of repeated complaints to Harcourt and MRC.

In summary, one must say that the administration of the program still showed many inadequacies. Perhaps this is inevitable in any experimental program carried out by people, many of whom are inexperienced at this sort of thing. However, the data analysis which was carried out greatly exceeds the typical analysis of such data, including as it did certain novel features such as the coordination of the "IN" and "OUT" cards with the test information, the testing of a representative state sample to provide a base for comparison of gains for the Title I children over seven months, and other things that will become evident as this report is completed. Care has been taken to point out these positive aspects of the program as well as the negative ones and the improvements over 1968-69 have been noteworthy.

During the summer of 1970, it was decided that all responsibility for statewide testing after that time would be transferred from Title I to the Director of Research and Testing of the State Department of Education. This decision was fully carried out and Title I did not participate at all in the collection of statewide testing data in 1970-71.

Two years of testing with Stanford, Form X, provided as much data-base information as was needed or desirable so far as Title I is concerned. The continuation of the use of Form X of Stanford in the upper grades in the fall of 1970, using over once again the same booklets that had been stored in the schools for the three-year period, was protested but economics again prevailed and the booklets were reused in spite of clear evidence of coaching in the 1969-70 program. The program was carried out through the auspices of the Bureau of Educational Research and Testing Services of UNH and the results are generally unknown to this writer. There was every reason to believe that this coaching would be accentuated through use of the same form a third year. Thus, Title I does not have the types of data reported in this document for the '70-'71 school year, except that "IN" and "OUT" cards are available which will reveal the extent to which the type of projects carried out are similar both in kind and proportion to what they were in '69-'70.

SECTION III

Part A

Chronological Age Distribution in New Hampshire Compared with the National Group

may not seem significant at first glance, the distribution of chronological ages within a defined group is actually very important in the interpretation of test data. Only the inexperienced or careless analyst forgets this. An older group (or child) will generally do better than a younger one even if the difference in age is only two or three months. For example, a child entering school about as late as he can enter and still be within the age-in-grade group, i.e., within the range of twelve months specified by local law or regulation as "normal", has a definite advantage over the youngest child admitted that year. This is just an offshoot of the basic fact that cognitive abilities as measured by mental ability or intelligence tests do contribute a great deal to the in-school performance of children, either separately or when considered as a group. Contrary trends are found in this area. Many communities are making day care and nursery school attendance an authorized and official part of public school instruction. Head Start is emphasizing structured pre-school experience for the disadvantaged. Other instances could be quoted.

Upward modification of the lawful entrance age to kindergarten or first grade is also being advocated and is being implemented in New Hampshire. This will have serious repercussions years later when the group as a whole, (if the practice becomes accepted generally in the state) becomes older than the national norm group, especially in light of the slightly above normal range of brightness now shown to be characteristic of our state population.

In this writer's opinion, there is no virtue of added age as a prerequisite to school entrance unless the home environment is adding something besides typical days-of-life experiences to the child's "readiness" for school which seems unlikely.

In Tables IIIA-1 and 2, we find the distribution of ages for boys and girls separately and for the total state groups in Grades 4 and 6 tested in the Fall of 1969. (The total group includes all those who did not code sex on the Otis-Lennon answer sheet from which these data came.) For no sensible reason that this author can discover, the coding of sex was omitted on many of the Otis-Lennon answer sheets.

We find that the New Hampshire total group appears to be a month younger than the group on which the Analysis of Learning Potential was

Table IIIA - 1

New Hampshire Statewide Testing Program - Fall 1969
Distribution of Chronological Ages
Separately by Sex and for the Total Group

By Way Of Comparison

Grade 4

Age in Years & Months	Boys	Girls	Statewide Total Group I*	Statewide Total Group II**	Nationwide ALP Norm Group Fall 1967
14- 4 to 14- 9	1	0	1	1	2
13-10 to 14- 3	0	1	1	1	1
13- 4 to 13- 9	3	1	4	5	12
12-10 to 13- 3	2	5	7	7	12
12- 4 to 12- 9	4	5	9	11	34
11-10 to 12- 3	16	13	29	36	70
11- 4 to 11- 9	89	38	127	142	162
10-10 to 11- 3	183	88	271	288	346
10- 4 to 10- 9	575	292	867	936	1,032
10- 3	132	81	213	224	215
10- 2	124	82	206	225	246
10- 1	149	91	240	256	303
10- 0	181	86	267	285	456
9-11	162	115	277	295	558
9-10	231	155	386	404	635
9- 9	342	356	698	729	830
9- 8	349	388	737	780	795
9- 7	394	390	784	839	890
9- 6	350	361	711	747	820
9- 5	362	393	755	786	871
9- 4	353	392	745	783	763
9- 3	336	408	744	772	835
9- 2	354	400	754	802	785
9- 1	330	427	757	802	756
9- 0	307	379	686	716	611
8-11	264	307	571	599	441
8-10	191	246	437	469	317
8- 4 to 8- 9	10	26	36	37	131
7-10 to 8- 3	20	24	44	46	47
7- 2 to 7- 9	1	0	1	1	0
Total N-Grade 4	5,815	5,550	11,365	12,024	12,976
N Mid-12 Months	3,972	4,447	8,379	8,824	9,149
N Mid-18 Months	4,911	5,057	9,968	10,513	11,127
% Mid-12 Months	68.30	80.12	73.73	73.39	70.51
% Mid-18 Months	84.45	91.12	87.71	87.43	85.75
Median Age	9-6	9-5	9-7	9-6	9-7

*Students Who Coded Sex
**Includes Students Who Did Not Code Sex

Table IIIA - 2

New Hampshire Statewide Testing Program - Fall 1969
Distribution of Chronological Ages
Separately by Sex and for the Total Group

By Way of Comparison

Grade 6

Age in Years & Months	Boys	Girls	Statewide Total Group I*	Statewide Total Group II**	Nationwide ALP Norm Group Fall 1967
17- 4 to 17- 9	1	0	1	1	0
16-10 to 17- 3	0	0	0	0	2
16- 4 to 16- 9	0	0	0	0	2
15-10 to 16- 3	0	0	0	2	3
15- 4 to 15- 9	1	0	1	1	12
14-10 to 15- 3	3	4	7	8	22
14- 4 to 14- 9	7	5	12	13	52
13-10 to 14- 3	19	11	30	32	79
13- 4 to 13- 9	116	46	162	174	229
12-10 to 13- 3	193	109	302	327	481
12- 4 to 12- 9	552	301	853	908	958
12- 3	127	59	186	200	226
12- 2	135	72	207	221	216
12- 1	119	82	201	214	309
12- 0	159	95	254	261	408
11-11	138	80	218	230	483
11-10	199	165	364	391	612
11- 9	347	357	704	737	803
11- 8	348	346	694	727	761
11- 7	376	402	778	816	809
11- 6	360	348	708	744	736
11- 5	385	378	763	800	735
11- 4	336	378	714	753	751
11- 3	308	386	694	726	855
11- 2	331	376	707	729	874
11- 1	316	411	727	766	754
11- 0	324	374	698	726	562
10-11	248	298	546	569	462
10-10	177	269	446	463	327
10- 4 to 10- 9	27	42	69	70	160
9-10 to 10- 3	46	44	90	93	73
9- 3 to 9- 9	1	0	1	1	0
Total N-Grade 6	5,699	5,438	11,137	11,703	12,756
N Mid-12 Months	3,878	4,323	8,179	8,556	8,735
N Mid-18 Months	4,733	4,876	9,609	10,073	10,683
% Mid-12 Months	68.05	79.50	73.44	73.11	68.48
% Mid-18 Months	83.05	89.67	86.28	86.07	83.75
Median Age	11-6	11-5	11-6	11-6	11-7

*Students Who Coded Sex
**Includes Students Who Did Not Code Sex

standardized.1/ The median age of the New Hampshire total group in Grade 4 in the 1969-70 program, with an "N" of 12,024, is 9 years and 6 months while the median for the ALP sample is 9 years 7. Both sets of data ostensibly were obtained in October of the school year although the spread of the testing dates in either sample is a factor of an unknown importance in this comparison. (This difference may be illusory because of slightly different times of year for the collection of the data.) For all practical purposes, we can say that the New Hampshire group now is fairly typical of the national sample as regards distribution of chronological ages at Grade 4. (National median, 9 years 7 months; New Hampshire, 9-6+.) The data for Grade 6 are consistent in this regard. The median age in Grade 6 is 11 years and 7 months in the Analysis of Learning Potential national sample and about one month lesa in the New Hampshire population.

It would hardly be fitting to leave this topic without commenting on the wide range of ages to be found within either Grade 4 or 6. Considering the statewide group in Grade 4, the effective ages range from 7 years 10 months (first age level for which there are a noticeable number of cases) to 12 years and 3 months (same limitation.) Thus the data show a real spread of more than four years. In Grade 6, the effective spread, 9 years 10 months to 14 years and 3 months, is, again, over four years. Thus retardation is clearly an accepted policy here, as it is generally, and its effects are cumulative from grade to grade. These over-age duller children have, nevertheless, increased in learning potential (mental power, not brightness) and thus are made more nearly equal by their retardation to their grade population peers in ability to handle the work of the grade. This is why the IQ is inappropriate as a basis for comparing capacity and achievement except for children within the age controlled range Young-bright and older-dull approach each other in ability to do school work.

To turn now to the age distributions separately by sex in Grade 4, it seems that girls in New Hampshire are a little over one month younger than boys. The same difference is found in Grade 6. This one month difference takes on more significance when we examine the results of the learning ability test (Otis-Lennon) discussed next in which it is evident that this measure of cognitive learning potential also favors the girls at both grade levels.

1/ The Analysis of Learning Potential was standardized in October 1967 and constituted the latest large and scientifically representative group available for comparison. The Metropolitan national sample of Fall 1969 confirms the ALP data but age data were not available in final form as shown here at the time of this report.

Perhaps the author may be excused if he does some further analysis of this largely overlooked problem. Many people see the increase in entrance-to-school age as primarily a logistics problem. In other words, it is reasoned that if the child is kept out of school until he is "ready" for kindergarten or first grade, he will be more likely to move through the grades at a steady pace, one year of school for one year of chronological age. Since this would substantially cut down on repeaters and thus get more children through school in the normal span of twelve years available for public education it would eventually, as the reasoning goes, save money for the taxpayer.

However, there is another way of looking at this problem, namely, the substantially large number of children who are above normal in their cognitive abilities who are unfairly denied the opportunity to enter school when they are capable of benefiting by instruction and even to move through school at an accelerated pace, if they are able. Such acceleration should not come about by double promotion but by widespread acceptance of individualized progress.

More and more the educational community is now seeing the school experience as necessarily being adjusted to the needs and capacities of the individual student. The idea that there is a fixed curriculum for a particular grade subject by subject, through which the individual proceeds in a kind of lock-step fashion, is totally fallacious. There is no generally accepted hierarchy of gradedness in the curriculum in any school subject, not even in arithmetic which, by the nature of the subject, might most closely approximate it.

There is a notion that there is a development age at which a child is "ready" for school and before which he is not. Back of this is the idea that some children are not ready to move out of the home environment into the broader group experience of the public schools at the normal time in their chronological age dimension. If there were a fixed curriculum, it would be true that some children would never reach the state where they could comfortably move out of the home circle into the larger school world on an equal footing. These are the educable and trainable children whose lot is certainly not an enviable one in today's schools.

Probably the truth of the matter is that the schools and the teachers in the schools are not ready rather than the children are not. It all depends on the essential goals of public education.

One could make a strong case for the idea that in a democracy, where education is publicly supported, it should be illegal to require that a child, not obviously a danger to himself or other children, be kept out of school when he has reached the mandatory school entering age. Parents might be advised to keep a child out of

school because, in the opinion of the school personnel, the child was not "ready" for the type of school experience a given school or system is ready to give. Along with such a recommendation should go a statement of philosophy stating the school's objectives and indicating that there is no intent or desire to individualize instruction. In other words, the curriculum for each grade is clearly set forth and the child must accomplish this curriculum in lock-step fashion if he is to be allowed to enter the group.

It is devoutly hoped that such a philosophy shall rapidly give way to a concept of education as "timely incremental learning" at a rate each child sets for himself and not at the rate the system dictates. At the present time, few schools accomplish this degree of freedom, although more and more are moving in this direction.

If one were to look at the problem solely from a cost and logistical point of view and be entirely consistent in doing so, the logical way to approach the problem would be to say that the public purse would support 12 years of school experience (or 10 or 14 as the case may be) after which the parents would have to assume responsibility for financing the child's education REGARDLESS OF HIS STATUS AT THE END OF THE TWELVE YEARS. This is now normally done after a student graduates from a senior high school, although there is growing awareness of the need for formal education through at least two more years.

The Age Controlled Sample and the Modal Age

Many years ago, Dr. Truman L. Kelley, then of Harvard University, developed the concept of the modal age. Dr. Kelley was interested in the problem of standardizing tests and wished somehow to reconcile the inconsistencies arising from the construction of age-oriented norms as compared to grade-oriented norms. To accomplish this, he created the idea of having children who were at age for grade used as the standard norm group. This would be a range of twelve months if all children whose birthdays fell within a calendar year would be allowed or required to enter school. This so-called modal age was to be determined by rather sophisticated statistical means but experimental evidence later showed that it could be more easily and almost as exactly determined simply by finding the range of twelve months of age containing the largest number of cases for any similar range in a given distribution.

Later, after this concept had been applied in the norming of the Stanford Achievement Test (1940 Edition), it was discovered that the range of twelve months was not adequate to take care of variations in entrance age and variations in promotional policies from one place to another around the country. Moreover the modal age population proved to be above average in measured intelligence. Subsequently, in the Metropolitan Achievement Test series the modal age idea was modified to include a range of eighteen months, thus allowing for variation in entrance age and also for differences in promotional policy. The 1958 edition of Metropolitan Achievement Test series was standardized on such an age controlled sample and the 1970 edition also will provide normative information for the age controlled sample.

This age controlled sample is an important concept because it provides an appropriate norm for the average child who has been allowed to move through school at the usual pace, i.e. one grade for each year of chronological age. For all children falling within the age controlled sample for their grade, one could say that the likelihood is great that his exposure to instruction had been more or less standard for a child of his age. For the older child who had been held back one or more years and thus is outside the age control group, two interpretations of score are necessary for complete understanding, one based upon his grade status regardless of his age and the other upon his age status regardless of his grade. This also should be done, of course, for the younger children who fell below the age controlled sample. These are the children who have been allowed to enter school at a younger than normal age or who have moved ahead of the group because of a more than average learning rate.

The age controlled sample provides a norm sample that remains comparable in both range of age and total in-school experience from grade to grade for 80% to 90% of children in school.

The age controlled sample in the distributions of chronological ages for the State of New Hampshire Grades 4 and 6 have been recorded, using a one month step interval for the requisite 18-montha age range. Note that this range is comparable to the age controlled sample for the country as a whole. About 87% of children in New Hampshire in Grade 4, for example, fall within the age controlled sample as compared to 86% in the national norm sample for the Analysis of Learning Potential. In Grade 6, the comparable percentages are 86% for New Hampshire and 84% according to the ALP data for the nation as a whole.

It would appear therefore, in conclusion, that in thia particular way of looking at the age data New Hampshire also is typical. The fact of thia typical character of the age composition of the New Hampshire population should be kept in mind later when this report deals with the achievement of New Hampshire students as compared with the national norm.

SECTION III

Part B

Evaluating the Measured Mental Ability of New Hampshire Students against National Norms

The Otis-Lennon Mental Ability Test was used in this program in both 1968 and 1969 with nearly identical results. There is great misunderstanding about the function of such tests as measures of cognitive learning ability. This ability is perhaps the most important dimension of school learning potential. Psychologists no longer consider it as a measure of native or inherited ability; it is a measure of the ability of an individual to respond to situations demanding the exercise of cognitive abilities, i.e. the ability to communicate, know, or understand and reason about people, places, or things.1/

In Table IIIB-1 we show the distributions of DIQs separately for boys and girls and for the total tested group in Grade 4 and in Grade 6. Please note that the sum of the cases tabulated separately for boys and girls does not always equal the total number of cases because a rather substantial number of pupils failed to code sex on their answer sheet and, therefore, could not be included in the distributions done separately by sex.

Table IIIB-1 accounts for roughly 12,000 children in each grade, a very large proportion of those in school in these grades. In Grade 4 the median DIQ is 101 and the mean 100.41, which is in itself one indication of the normality of this distribution. The standard deviation is 14.75. (In an unselected age population, including all children of a given age regardless of grade, the standard deviation would be 16.0 by definition.) One can see that the range of DIQs is from about 50 to 150, thus covering the total range of the test. The distribution of DIQs for Grade 6 also is given in Table IIIB-1. These data show that the situation does not differ very much from that in Grade 4. The median (50th percentile) DIQ is 102 and the mean (rounded off) is the same. The standard deviation is 15.12 and the range is from DIQ 50 to 150.

Similar distributions are available for Grade 2 and Grade 8. In Grade 2 the median DIQ is 103 and the mean also is 103 with a standard deviation of 14. In Grade 8 these values are 103 for the median, 104 for the mean with a standard deviation of 14. The consistency in these results over the four grades is notable. Distributions for Grades 2 and 8 are not shown here because the major part of this report is limited to Grades 4 and 6 for economy's sake.

1/ See "Intelligence", Encyclopedia Americana, 1971 Vol. 15 Pgs. 241-245

In these DIQ tables, we have added one additional feature, namely, cumulative percentages, which allows the reader to determine the percentage of children having DIQs below any desired point corresponding to the uppermost value in the step interval. For example, consider the step interval 111-113; 84% of the boys in Grade 4 had DIQs of 113 or lower. By contrast, in the interval 114-116, 85% of the girls in Grade 4 had DIQs of 116 or lower, clearly showing a differentiation between boys and girls in favor of girls.

The statistics given in this table are summarized at the end of the table where percentiles (scores) are given corresponding to selected percentile ranks together with the mean and the standard deviation of each distribution.

Means and medians generally agree in near-normal distributions and it will be seen there is such agreement in this instance. Skewness in a distribution will be reflected in a difference between these two averages, the mean always being toward the "tail" of the distribution from the median. (Skewness in lay terms might be described as lack of symmetry or lopsidedness.) If a test is too easy, for example, there is a tendency for cases to pile up at the top end of the score scale while the opposite is true if the test is too hard. This difficulty was encountered in certain of the distributions of subtest raw scores on Stanford.

The New Hampshire populations in Grades 4 and 6 as indicated by these distributions are so typical of the national scene in both age and learning ability that they might well have been used, with only minor loss in precision, to provide national norms for the Otis-Lennon Test! Even the standard deviations, which by definition for the national population are 16 for any single age group, closely approximate 15 in our grade distribution where a slight curtailment is usually found.

A realistic appraisal of what the demonstrated variability of these brightness measures mean for the New Hampshire educational program is perhaps the strongest argument for individualizing instruction, recognizing that many children must necessarily proceed through the established curriculum at a somewhat different pace than the average or typical child depending upon their ability to cope with the school learning situation.

It is for this very purpose that much Title I money is spent, namely, to help individualize instruction for needful children, especially those considered disadvantaged and/or those with known and definable weaknesses in learning. It is hoped of course that such special help will restore these pupils to their rightful place in the distribution of scores in the various traits measured. What actually happened among tested Title I children will be discussed later. However, there are few instances where low DIQ children ever have become average or high on these measured traits except where it has been possible to show that the original test was inappropriately

Table IIIB-1

New Hampshire Statewide Testing Program
Distribution of Otis-Lennon Deviation IQs
Separately for Boys, Girls, and Total Group
Tested Fall 1969

	GRADE 4						GRADE 6					
	BOYS		GIRLS		TOTAL*		BOYS		GIRLS		TOTAL*	
DIQ Interval	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age
150	5	99	4	99	9	99	5	99	7	99	12	99
147-149	2	99	2	99	4	99	7	99	2	99	9	99
144-146	2	99	5	99	7	99	13	99	10	99	25	99
141-143	10	99	10	99	21	99	23	99	25	99	50	99
138-140	19	99	21	99	41	99	21	99	13	99	35	99
135-137	19	99	30	99	54	99	24	99	22	99	47	99
132-134	38	99	48	98	92	99	42	98	55	98	101	98
129-131	54	98	73	98	129	98	64	98	88	97	157	98
126-128	93	97	77	96	177	97	84	97	94	96	181	96
123-125	97	96	130	95	240	95	197	95	247	94	457	95
120-122	120	94	189	92	330	93	168	92	166	89	349	91
117-119	194	92	236	89	446	90	260	89	250	86	533	88
114-116	250	89	293	85	570	87	292	84	336	82	652	83
111-113	357	84	457	79	840	82	312	79	386	75	726	77
108-110	325	78	397	71	760	75	367	73	412	68	818	71
105-107	416	73	454	64	920	69	388	67	439	61	880	64
102-104	526	65	549	56	1133	61	449	60	519	52	1003	56
99-101	420	56	462	46	921	51	439	52	466	43	959	48
96-98	419	49	376	38	841	44	468	44	414	34	929	39
93-95	465	42	365	31	880	37	439	36	420	26	906	31
90-92	366	34	344	24	756	29	367	28	276	19	676	24
87-89	455	27	286	18	796	23	313	21	200	13	554	18
84-86	321	19	208	13	565	16	210	16	123	10	354	13
81-83	206	14	141	9	376	12	164	12	104	7	284	10
78-80	177	10	111	6	321	8	139	9	74	6	233	7
75-77	135	7	79	4	230	6	91	7	59	4	160	5
72-74	92	5	53	3	156	4	78	5	39	3	124	4
69-71	65	3	46	2	118	2	58	4	34	2	94	3
66-68	48	2	24	1	75	2	44	3	38	2	85	2
63-65	33	1	9	1	45	1	37	2	24	1	61	1
60-62	12	1	12	1	25	1	14	1	11	1	26	1
57-59	15	1	7	1	24	1	14	1	9	1	26	1
54-56	7	1	5	1	14	1	9	1	4	1	13	1
51-53	8	1	3	1	12	1	8	1	3	1	11	1
48-50	5	1	4	1	10	1	17	1	2	1	19	1
Totals	5,776		5,510		11,938*		5,625		5,371		11,549*	
Q3 %ile 75	108		111		110		111		113		111	
Q2 %ile 50	98		102		101		100		103		102	
Q1 %ile 25	88		93		90		91		95		92	
Mean	98.88		102.32		100.41		101.01		103.91		102.32	
Standard Dev.	14.94		14.27		14.75		15.68		14.45		15.12	

*The Distributions of DIQs for Boys and Girls do not sum to the Total Distribution because of the failure of a substantial number of pupils to code sex.

administered. An instance would be the administration of the test to a nearly non-English speaking child with severe listening and reading problems.

The major reason for computing a measure of brightness is to determine to what extent we can expect above or below normal achievement for any child, assuming that he is of normal age for his grade placement and has had a more or less normal experience in school, i.e., has not been absent extensively, for example. A much better way of making systematic comparisons of capacity and achievement is to use grade based norms. Such norms are available for the Otis-Lennon both as national and as local (state) stanines, and for each of the Stanford Achievement Tests.

Raw score distributions on the Otis-Lennon test also are available. The median raw score for Grade 4, for example, is 32.9 out of a possible 80 and the mean is 34.0. Both of these averages would put the children in New Hampshire within the fifth stanine nationally, i.e., within the normal range for these grades. More precisely, the Grade 4 median raw score would have a percentile rank of 54 and the median raw score of Grade 6 would have a percentile rank of 55.

Local vs National Norms

Stanines based on score distributions for a local population such as a single grade in this state, a school district, or even a school, are better than national norms when one wants to cancel out systematic population differences in achievement from subject to subject when comparing capacity and achievement. State and local stanines based on raw scores were made available for Stanford and Otis-Lennon in both 1968 and 1969. As the next part of this study, bivariate (two-way) distributions were made for Otis-Lennon raw score stanines versus each of the tests in the Stanford Battery given in the fall.

It would be much too space-consuming to reproduce all bivariate charts for the Otis-Lennon Mental Ability Test stanines versus all of the subtests in Stanford for both grades, but one such chart has been reproduced from the 1968 program in order to illustrate several points which are very important concerning the technique for and contribution of the bivariate distribution as a way of comparing school learning capacity as measured by a mental ability test with measured achievement in school.1/

Selected for this purpose was the Otis-Lennon Mental Ability Test state raw score stanines versus similar Stanford Paragraph Meaning stanines in Grade 4. A sentence or two about stanines may be important here for the general reader. Stanines are, essentially, normalized standard scores, which means that they have the characteristics of an equal-unit scale. For example, the rungs of a ladder are equally spaced apart and thus may be considered, in a sense, an equal-unit scale. In a somewhat analogous sense stanines are like 9-step ladders having rungs equally spaced.

Perhaps we can look at this bivariate distribution without getting more deeply into the complications of how stanines are computed at this time.

See Chart IIIB-I.

It is easy to see that there is a general drift in the cell frequencies from lower left to upper right with a concentration of cases appearing along the mid-diagonal line, shown by the dotted line. If we mark off one stanine cell at each stanine level to the right and left of this mid-diagonal by zigzag lines, we will have the mid-stanine band or range. On Chart IIIB-I a large proportion of the cases in the distribution is found within this band. As a matter of fact, the percentage of cases falling within this mid-stanine range is virtually of the same magnitude as the Pearson product moment correlation coefficient as reported, which is .73.

All test scores are subject to some kind of measurement error, that is, variation due to chance factors that cannot be identified or controlled. Usually one stanine accounts for better than one standard error of measurement expressed in raw scores. Thus we can say, for all practical purposes, that most of the youngsters falling within this mid-stanine range are indeed performing in Paragraph Meaning in a manner consistent with their mental ability stanine as measured by the Otis-Lennon. Please remember that these stanines are not based on DIQs but are based upon the distributions of raw scores. All fourth graders tested in the State of New Hampshire in Fall 1968 are included. Thus, roughly one-quarter of the students fall outside this mid-stanine band for many reasons, about half above and half below the band.

A scattering of cases show a surprising contrast between performance on the mental ability test and performance on the Paragraph Meaning. For example, there is one child shown who had a stanine of 9 on the Otis-Lennon Mental Ability Test but only a stanine of 1 on Stanford Paragraph Meaning. Such deviant individuals are very rare and almost always can be accounted for by some digression from good testing practice or by actual errors in taking the test or in data processing. It is standard operating procedure and a very highly recommended practice, that students falling outside the mid-stanine range be studied more carefully than those within this range to be sure there are no such irrelevant factors involved. If such factors can be identified, corrective action can be instituted in one way or another. This would be especially true of the very extreme cases noted.

1/ 1969 data were unavailable to the writer at the time this report was prepared.

Chart IIIB-1

A Representative Bivariate Chart or Correlation Plot Showing Graphically the Degree of Correspondence Between Paired Scores*

Stanford Paragraph Meaning

Otis-Lennon	1	2	3	4	5	6	7	8	9	Stanine ←
9	1			1	7	18	96	165	287	575
8		1	1	11	37	97	291	236	115	789
7	6	8	19	72	213	416	620	305	65	1724
6	19	44	86	291	638	668	513	137	22	2415
5	40	124	254	579	853	513	249	18	1	2631
4	95	233	376	696	596	223	47	6		2272
3	105	296	372	470	259	67	24	3		1596
2	115	257	293	250	116	21	4	1	2	1059
1	115	123	107	112	30	7	2			496
Stanine ↑	496	1086	1508	2482	2749	2030	1843	871	492	13557

r = .73

% in Mid-Stanine Band = .73

*From the Statewide Testing Program for Grade 4 Fall 1968; stanines, in both cases being based on raw scores.

Table IIIB-2

New Hampshire Statewide Testing Program - Fall 1968
Correlations - Otis-Lennon and Stanford Tests

	Grade 4		Grade 6	
Test	Stanine	Raw Score	Stanine	Raw Score
Word Meaning	.73	.76	.74	.74
Paragraph Meaning	.73	.76	.77	.76
Language	.74	.75	.78	.78
Spelling	.63	.66	.61	.61
Word Study Skills	.68	.70		
Arith. Computation	.42	.44	.51	.51
Arith. Concepts	.64	.68	.69	.68
Arith. Applications	.66	.69	.71	.72
Social Studies	.72	.74	.76	.75
Science	.73	.76	.73	.73

If a test result is suspicious in terms of the student's previous performance, the recommended procedure would certainly be to retest that individual. Both of the tests being compared have very short time limits indeed compared to the many hours spent learning to read. Surely, it is not fair to a child to judge him on the basis of any one pair of such test scores. Therefore repeated testing, preferably cumulative over a period of years, is highly recommended. Obviously, one mental ability test in the mid-elementary grades is entirely insufficient. In many casea using one test result, not substantiated by others, can do irreparable damage to individual pupils. This is especially true of those having greatly deviant scores between normally highly correlated tests. This disadvantage is maximized if teachers, parents or children accept such a single test result as definitive and final. GOD FORBID!

Relationship of Measured Mental Ability with Measured Achievement

Just previously, we have seen what the bivariate distribution or correlation plot looks like when measured mental ability (Otis-Lennon) is compared with a specific measure of a curriculum oriented variable, namely Paragraph Meaning. The product moment correlation was .73 for this particular chart. (Note that when correlations for these same data were computed in terms of raw scores, the values reported from the computation center were slightly higher. Neither one is "wrong"; using ungrouped raw score data with no real operational limits on the precision of the computations in terms of decimal fractions retained, etc., just yields a slight but insignificantly higher value. The coarseness of grouping involved in making the stanine bivariates is a factor, but this slight difference in the r's is a small price to pay for the advantages of seeing exactly what the correlation plot looks like.)

In Table IIIB-2 all remaining correlations of Otis-Lennon with Stanford Achievement subtests are shown for both raw scores and stanines and for Grades 4 and 6.

In such a table Arithmetic Computation almost always is lowest with Spelling usually next in order. This is due to the tendency to teach these subjects in a more nearly rote fashion. Reading ability and reasoning ability are less important in these areas than in the other subjects. Even in communities having a well established modern math curriculum these correlations will be low because the tests emphasize outcomes rather than process.

It is here that the finding that the percent of cases in the mid-stanine band closely approximates the stanine chart correlation coefficient becomes really important to understanding what this table means. The coefficient subtracted from 1.00 gives the percent outside the band. When this value is split in half, roughly one half can considered above the diagonal band and one half below. NOTE: not all cases requiring special attention are OUTSIDE the band; only those where there is a serious inconsistency in paired test results.

Some youngsters may be in such trouble that all objective measures uniformly underestimate their real school learning potential and real achievement potential. A spastic child or one having visual problems will do poorly on objective tests of ALL kinds. Far too often such children will also be under-rated by their teachers. It is at this point that one can see most clearly the need for a high level of competence among school teachers in understanding and using test results AND ALL OTHER OBJECTIVE EVALUATION DATA.

SECTION III

Part C

Overall Description of the New Hampshire Population as Regards Achievement

It is very pertinent to ask what kind of achievement is characteristic of children attending the public schools of New Hampshire in terms of national norms on the Stanford Achievement Test. In the 1969-70 testing program, exactly the same achievement tests were administered as in 1968-69, namely Form X of Stanford at all tested grades (except Primary II, Form W in Grade 2 in the spring of 1969).

The data presented herein are not inconsistent with the data for 1968-69 but, since strenuous efforts were made to improve the administration of the tests in 1969-70 and since available data for spring and fall testing of Title I cases is limited to 1969-70, only the 1969 Fall testing program data will be considered in this section.

In Table III-C-1 the raw scores corresponding to selected percentile ranks 75, 50 and 25 are tabled for each of the tests in the Stanford Battery. The raw score percentiles were then expressed in terms of Stanford grade equivalents for both Grade 4 and Grade 6. These norms are based on about 98% of the norm group tested in March 1963, 1% to 2% being eliminated as being extremely atypical as to age. These Stanford data are shown to the left of the vertical line separating the Stanford information from the derived Metropolitan normative information.

The New Hampshire norm for Grade 4 would be a grade equivalent of 4.2 and, similarly, the norm for Grade 6 would be 6.2 since the tests were administered in October.

An examination of this table shows that the raw score percentiles, when transformed into grade scores or grade equivalents, are below the Stanford national norms on every test in Grade 4 and also in Grade 6. This is true in spite of the

fact that the State is above the national norm on the Otis-Lennon Mental Ability Test at these grade levels.

However, an examination of the Stanford Norms booklet and the Technical Supplement reveals that the grade populations used for norming the Stanford series, while including very large numbers of cases, were above average in brightness, the deviation IQs on the Otis Quick-Scoring Mental Ability Test being as follows: Grade 2, 105, Grade 4, 109, Grade 6, 109, and Grade 8, 108. No one knows why this upward deviation occurred. Unusual care was used in selecting the norm sample when this series was standardized but some unknown bias resulted in unrepresentative samples as regards brightness. This in turn seems to have resulted in norms that were too "hard". "Hard" is in quotes because it is evident from an examination of the relationship between total possible score and raw score corresponding to the selected percentile ranks that the tests actually were on the easy side. In this context "hard" means that a lower grade equivalent was assigned to each score than subsequent experience seemed to justify.

The aim is to have the norm fall in the middle of the range of possible scores in order to measure all achievement levels in the tested group. Stanford seems to have met this criteria in Grades 4, 6 and 8; the Primary I Battery, however, was much too easy, if one may judge on the basis of the raw score distributions.

Furthermore, Stanford was standardized in March while our group was tested in October. Spring norms generally run "harder" than fall norms for reasons not too well understood and too complicated to explain here.

The important question is whether New Hampshire children are really achieving as poorly as Stanford norms indicate. Some light can be shed on this matter by considering New Hampshire achievement in terms of the more up-to-date Metropolitan '70 norm data. This is made possible by the publication of equivalence tables for Stanford and Metropolitan '70.

The revised Metropolitan Achievement test was standardized in 1970 on a national stratified random population. During the period between the standardization of the Stanford in 1963 and Metropolitan in 1970, substantial developments occurred in the technology of choosing stratified random samples for normative purposes.1/ Also, great strides were made in computer technology and capability. Thus, the new Metropolitan norms must be considered intrinsically superior to, i.e., more representative than, the Stanford norms. Moreover no one can discount the fact that important curriculum changes did occur during this period.

In the 1958 edition of Metropolitan the widely used norms were "age controlled" norms, i.e., those children most likely to be at grade for age. The published Metropolitan '70 revision presently provides only norms for total population with no elimination of over-age pupils. Since Stanford norms used nearly all cases, regardless of age, these are more comparable in range of age than age controlled norms. Age controlled norms also will be generally available for Metro '70 shortly and in the writer's opinion are much more serviceable at a time when there is growing emphasis on individualization of instruction.

It is routine procedure at Harcourt Brace Jovanovich, whenever either the Metropolitan or the Stanford series is revised, to carry out a careful series-to-series equating program in order to facilitate going from one series to the other for those who wish to evaluate differences from series to series which may be due to norm differences. It was the existence of such tables of equivalence, given in grade equivalent terms, that made the above described comparisons possible in this study.

Look now at the equivalent Metro '70 grade equivalents shown in the column to the right of the Stanford values. The net effect is to suggest that New Hampshire is, in truth, performing more nearly up to its measured capacity than indicated on the basis of Stanford norms. For the most part, this section of the table strongly indicates that there should be no real concern for the level of educational achievement in this state. The low points at Grade 4, according to Metropolitan norms, are Social Studies and Science where the deficit amounts to .2 of a calendar year at Grade 4 but is at the norm in Grade 6. In Grade 6, performance in Arithmetic Computation and Arithmetic Concepts shows a deficit of .2 of a year. In all other tests the state group was at or above the Metropolitan total population national norm.

Literal interpretation of these equating tables does involve some risk of over-simplification. For instance, the Stanford Paragraph Meaning Test may measure slightly different aspects of reading than the Metropolitan Reading Test. Correlational data are lacking at this moment, but the writer's experience over the years has been that Stanford and Metropolitan tests correlate about as well as two forms of either battery, especially for the basic skill subjects. It is also noteworthy that the technique used for standardizing the Otis-Lennon Test, which is a relatively new test, was substantially similar to that used in standardizing the Metropolitan; in fact, Otis-Lennon was the precursor of the Metropolitan procedure in most basic aspects.

1/ Perhaps the most notable study in this area is that conducted by Dr. Thomas P. Hogan entitled, "Socioeconomic Community Variables as Predictors of Test Performance". Dr. Hogan also was responsible for selecting the normative sample used in the Metropolitan standardization program.

Summary

From the data in Table III-C-1 it would seem evident that we can consider New Hampshire a very typical state indeed in terms of Metropolitan national norms. Much of the concern expressed at the State Department level and in the communities throughout the state over the poor New Hampshire showing appears to be attributable to lack of representativeness in the Stanford norm population, plus some curricular changes, especially in arithmetic. There is evidence in the tables equating the '58 and '70 Metro editions that arithmetic achievement has declined, especially in Computation. The writer is inclined to attribute the shift (if it really happened) to the haphazard, unsystematic way that modern math was introduced in the schools, plus a generally agreed lack of sufficient maintenance-of-skills work in the new math.

The generalization made about average (median) performance applies with almost equal force to the 25th and 75th percentiles, also expressed as grade equivalents.

A word of caution is necessary. Grade equivalents are not ever equal units from test-to-test within the same series or from level-to-level within the same test. Reading grade equivalents are NOT comparable to Computation grade equivalents. Standard deviations of grade equivalents tend to vary inversely with the gain in score points from grade-to-grade. For example, Language generally has the smallest increment in score and the largest standard deviation in grade equivalents. Computation, generally, has a large score increment from grade to grade and thus the smallest standard deviation of grade equivalents. For this reason, pupil profiles in grade scores or grade equivalents are essentially meaningless.

The grade equivalent, as usually computed, is based on the assumption of continuous development over a calendar year's time since a grade equivalent norm line is drawn through the median scores for successive grades, i.e., for children differing in age and life experience by a full calendar year. In point of fact it is a trend line through the average scores of successive grades tested at the same time in the school year. This ignores differences that may occur subject-to-subject due to differential growth (or forgetting) during the summer vacation. The effect of summer forgetting or non-learning can be very great. Arithmetic computation skill, for example, is rarely learned to any degree out of the specific instructional environment of the school, whereas reading and vocabulary continue to develop due to general life experience and mental development. In spite of this obvious 12-month span, grade equivalent points, derived by dividing the total gain in raw score from grade-to-grade into ten parts, are often erroneously called months. Since 180 days of schooling is about maximum for most systems, even dividing by 10 would be a doubtful procedure if the resulting units are to be called "months"; nine months is more representative of the amount of time between opening and closing of school, especially if within-school-year vacation time is taken into account.

Profiles are sensibly plotted only in terms of grade-based standard scores with some semblance of equality of units over the scale range. Use of such standard scores makes peer comparisons comparable from test-to-test within a grade level and generally for the same test at successive levels. Stanines are by all means the most useful of such standard scores. National stanines are now provided routinely but State stanines are preferable for pupil profiles. Such, as will be shown later, have been provided in New Hampshire.

Table III-C-2

Although most of this analysis will be concerned with interpreted scores, grade equivalents, percentiles, stanines (local and national) and the like, it is important to examine the raw score characteristics of each of the tests used at the two grades chosen for this study, namely Grades 4 and 6. An examination of these data reveals characteristics of the tests in a way that cannot be seen by any other means of analysis.

A good test, in the technical sense, for a particular population would be one where the average score (either mean or median) is far enough above a zero score and far enough below a perfect score to permit all the individuals tested to indicate what they are capable of doing. For example, a median raw score of 15 out of 38 points on Word Meaning for Grade 4 is marginal. The test appears to be too hard. On the other hand, there are only 38 items so the other possibility is that the test is too short for separate interpretation. Certainly the bottom 50% of the group cannot be very well distributed with only a total of 14 additional points of score available to represent all levels of achievement from lowest to the average.

Generally speaking, the Intermediate II Battery used at Grade 6 does a better job in this respect than the Intermediate I at Grade 4. Here there are 48 items in the Word Meaning Test and the median for the state is 25; the 75th percentile is 32, leaving plenty of "top" while the 25th percentile of 19 leaves much more "bottom" to take care of the less able youngsters.

A comparison of the means and the medians for each test, given in parallel columns, will indicate to some extent the skewness in the distributions.

Recall from our previous discussion that in a skewed distribution the mean is always in the direction of the long tail as compared to the median. None of the tests tabled here seems to be seriously skewed, but an examination of the complete distributions of scores that were made, test by test, in the entire state in 1968-69 and again

Table III-C-1

EQUIVALENT GRADE SCORES
In Terms of Metropolitan '70 Norms
For 25th, 50th and 75th Percentile Ranks
New Hampshire Statewide Testing Program
October 1969
Statewide

GRADE 4					GRADE 6				
	Stanford Int. I			MAT '70		Stanford Int. II			MAT '70
Test 1/	%ile Rank	Raw Score	Grade Equiv.	G.E.3/	Test	%ile Rank	Raw Score	Grade Equiv.	G.E.3/
(38) Word Meaning	75	19	4.6	4.9	(48) Word Meaning	75	31	7.1	7.8
	50	14	3.8	4.1		50	24	5.9	6.4
	25	9	3.2	3.5		25	18	4.9	5.2
(60) Paragraph Meaning	75	29	4.4	5.0	(64) Paragraph Meaning	75	40	6.7	7.6
	50	22	3.7	4.2		50	31	5.6	6.3
	25	16	2.9	3.2		25	23	4.6	5.3
(50) Spelling	75	28	4.6	5.0	(56) Spelling	75	36	7.0	7.3
	50	19	3.8	4.1		50	28	5.9	6.2
	25	13	3.2	3.4		25	20	5.4	5.7
(61) Word Study Skills 2/	75	44	5.5		NO TEST				
	50	34	3.9						
	25	24	2.7						
(122) Language	75	73	4.5	5.3	(134) Language	75	95	7.1	7.9
	50	62	3.5	4.3		50	82	5.7	6.4
	25	52	1.7	1.7		25	70	4.5	5.3
(39) Arithmetic Comp.	75	14	4.0	4.6	(39) Arithmetic Comp.	75	17	5.9	7.2
	50	11	3.6	4.1		50	13	5.2	6.0
	25	8	3.1	3.5		25	9	4.4	5.1
(32) Arithmetic Concepts	75	16	4.8	5.2	(32) Arithmetic Concepts	75	16	6.5	7.1
	50	11-12	4.0	4.3		50	12	5.6	6.0
	25	8	3.0	3.2		25	9	4.9	4.6
(33) Arithmetic Appl.	75	16	4.6	5.0	(39) Arithmetic Appl.	75	22	6.6	7.4
	50	12	4.0	4.2		50	16	5.6	6.3
	25	8	3.4	3.6		25	11-12	4.5	4.9
(49) Social Studies	75	25	4.5	4.7	(74) Social Studies	75	47	6.8	7.8
	50	20	4.0	4.0		50	37	5.6	6.2
	25	15	3.5	3.5		25	29	4.8	5.2
(56) Science	75	31	4.6	5.0	(58) Science	75	37	6.7	7.7
	50	23	3.9	4.0		50	30	5.6	6.2
	25	17	3.5	3.5		25	23	4.4	4.7

1/The number in the ( ) is the number of items on the test.

2/No comparable test in Metropolitan '70 Elementary or Intermediate I Battery

3/These are Metropolitan Total Population norms, which are most nearly comparable to Stanford.

Table III-C-2

Analysis of Test Characteristics Relating to "Goodness of Fit"
Of each Test for the Level at which It Was Used
For the Total Population Tested and for the Random Sample

Grades 4 and 6 - Statewide Testing Program Fall 1969

STANFORD ACHIEVEMENT TEST

Intermediate I Battery — Intermediate II Battery

	GRADE 4						GRADE 6					
			Total Population		Random Sample				Total Population		Random Sample	
Test	No.of Items	%ile Rank	Select. %iles	Mean	Select. %iles	Mean	No.of Items	%ile Rank	Select. %iles	Mean	Select. %iles	Mean
Word Meaning	38	75	20.0		21.0		48	75	31.5		32.0	
		50	15.0	15.0	15.5	15.9		50	25.0	25.0	26.0	25.7
		25	10.0		10.5			25	18.5		19.5	
Para. Meaning	60	75	29.0		31.0		64	75	41.0		41.0	
		50	22.0	23.4	23.0	24.4		50	31.5	32.2	33.0	32.9
		25	16.5		17.5			25	23.5		25.0	
Arith. Comp.	39	75	14.5		14.5		39	75	17.0		17.0	
		50	11.0	11.6	11.0	11.5		50	13.0	13.6	13.5	13.9
		25	8.5		8.5			25	9.5		10.0	
Arith. Conc.	32	75	16.5		16.5		32	75	17.0		17.0	
		50	12.0	12.7	12.0	12.9		50	13.0	13.3	13.0	13.5
		25	9.0		9.0			25	9.5		9.5	
Arith. Appl.	33	75	17.5		17.5		39	75	22.5		22.5	
		50	12.5	12.7	12.5	12.9		50	16.5	17.4	16.5	17.6
		25	9.0		9.0			25	12.0		12.0	

OTIS-LENNON MENTAL ABILITY TEST

Elementary II Battery

Otis-L Raw Score	80	75	43.1				80	75	65.0			
		50	32.9	34.0				50	54.7	52.4		
		25	24.0					25	42.3			
Otis-L IQ		75	110.5		111.5			75	112.0		112.0	
		50	100.5	100.4	102.5	100.8		50	102.0	102.3	102.5	102.4
		25	90.0		93.0			25	93.0		93.0	

in 1969-70 reveals that the distributions tend to be toward the low end of the scale. The distributions are not noticeably skewed however.

Data for the Otis-Lennon Mental Ability Test is shown at the bottom of the page. This test was administered only in the fall while the Stanford Tests were repeated in the spring with selected samples. The Otis-Lennon was a revision of the Otis Quick-Scoring series, improved to provide a better measure of "G" or general factor of intelligence which is a kind of overriding mental ability which, taken in toto, seems to correlate fairly highly with various criterion measures. It yields a single measure of brightness called a deviation intelligence quotient (DIQ). It also has grade oriented norms (percentile ranks and stanines) which greatly enhances its usefulness. Of the tests in the 1969-70 statewide battery, Otis-Lennon does a better job of predicting success in any subject matter field than does any other test. (See the intercorrelation table in Appendix.) The Otis-Lennon stanine, for this reason, is given a weight of 3 when included in the composite prognostic score compared to a 2 as the highest weight given any other single test. 1/

Mental ability tests, such as Otis-Lennon, have been criticized by many people who lack fundamental knowledge of mental measurement as being unfair for many children from an environmental point of view. The same argument, of course, can be made for a reading test or a spelling test or any other kind of test if one realizes that performance in any of these areas is not totally determined within the bounds of the instructional program in the school. Mental ability tests have a tremendous usefulness for teachers who understand their strengths and limitations since they provide another look at the child in the broad spectrum of his intellectual or cognitive development. Anyone who has studied a correlation matrix such as that mentioned above, especially if he examines the actual bivariate charts, can't reasonably be worried about determinism, i.e., the self fulfillment prophecy.

From the raw score data, it is apparent that this test is functioning very well since both the median and the mean are substantially above the chance level at both grades. The mid-value between the median of Grade 4 and Grade 6 would be a score of 44 out of 80 possible. (Note that the same test is used at both grade levels.) This represents about the best possible fit that one could obtain for a test designed to be used over a three-grade range, as this one is.

1/ This does not mean that there are no coefficients higher than Otis-Lennon. Two arithmetic tests may intercorrelate higher than Otis-Lennon with either, but will not correlate as high with other subject tests. Perhaps it is better to say that the median of the Otis-Lennon correlation with Stanford subtests is higher than a similar statistic for any other test.

SECTION III

Part D

Variation Among Communities

In this Title I report we would like to be concerned with the performance of individual children within each school district as the most significant breakdown in the report. Hopefully, in many instances we might investigate the performance of Title I children within individual schools. However, the total number of cases available in Title I in any one school almost always is too small to make any valid statistical comparisons within a school, although much more could be done if school and district distributions were conveniently available. Thus one is left with the comparison of results from Fall to Spring for individual pupils against all Title I cases as a reference population and/or with the random sample.

However, it was possible to obtain school district means separately by test for the 137 districts making up the tested state population. Since this information was in raw score form, the resulting means are not comparable from test to test. To meet this need, a system of standard scores was inaugurated by making a linear transformation of the raw scores for the total state on the basis of the pupil score distributions so that the mean score for every test administered at each of these grades (Grades 2, 4, 6 and 8) was assigned a value of 50 and a standard deviation of 10.

Note these were pupil distributions. Transformation tables were generated by computer, were printed and made available for any who needed or wanted to have them.

The description of the New Hampshire situation would be incomplete without studying the variability among school districts. To accomplish this and to facilitate making school district (or school) profiles possible for satellite studies the raw score means of each school and school district were transformed using the new linear standardized scores. Listings were made by school and school district and were turned over to the Department of Education for appropriate use.

In this portion of the report we present graphically the distributions for 137 school districts in the state in terms of these pupil-based linear standard scores.

One not aware of the reality of community differences cannot help but be astounded at the spread of these standard scores on all tests, including the Otis-Lennon. One would expect the range of standard scores for district means to be much less than it was, although the mean of these distributions should approximate the mean of the population, as it did. These distributions are shown in Table III-D-1 and Table III-D-2 sepa-

rately by grade and subtest together with an effective graphic display (histogram) of the distributions. 1/

The value of each * varies from figure to figure. The mode ALWAYS is set equal to 50, i.e., is represented by 5 *. The value of the * for any particular figure can be obtained by dividing 50 by the number of cases having the modal score. For example, in Table III-D-1 showing the distribution of Otis-Lennon standard scores the modal number is 26 so 50/26 or .52 is the value of a single *.

This procedure has the virtue of making the shape of the distributions visually comparable from one histogram to another and also compensating for wide swings in the number of cases. For example, it serves as well for statewide pupil distributions of 13,000+ cases as for these distributions of 137 districts.

In the Grade 4 distributions, the range for Otis-Lennon standard scores actually is from a standard score of 31 to 61. Discounting the one extremely low standard score, namely 31, the range still is from 36 to 61 with more or less a continuous distribution between these points. The computed standard deviation of 4.28 even more strikingly indicates this variability among districts since this approaches one-half of the assigned standard deviation of 10 for pupil scores. The mean of 49.28 for Otis-Lennon is perhaps unduly influenced, in view of the relatively small number of school districts, by the one extreme case where the standard score assigned was only 31. Even so, it is not far off from the established mean of 50.

Looking now at the Otis-Lennon distribution for Grade 6, with the same 137 school districts involved, we find one school district with a mean of 24, another one with a mean of 30, after which there is a skip to 41. The two extreme values must be discounted, but even doing this the range from 41 to 57 is very substantial indeed. Here again the standard deviation of 4.21 approaches half the standard deviation of the total pupil distribution and the mean of 49.5 is approximately that of the population.

One cannot help but ask how it can happen that entire school districts (granting that in some cases the number of children tested still is quite small) can show this kind of variability. Without a careful study of data not currently available to this writer, of such factors as socio-economic status within the community, amount of money spent for education, the quality of instruction and of instructional materials and, finally, the extent to which other unspecified factors are influencing the data, no explanation is possible. One only has to face the hard reality that this is so and that there is little or no chance that this is due to anything other than factors quite independent of the test instruments used or any other factor that is related to the testing program. The data, in this writer's opinion, describes accurately existing conditions within the school districts of this state.

Perhaps the best confirmation that this variability of district means on Otis-Lennon is no chance finding lies in the fact that the distributions of achievement test means for these same communities show the same phenomena of wide variability in community means translated to comparable terms by means of this rescaling technique. The writer has chosen to use the mental ability test for discussion purposes because it seemed advisable to escape the trap of saying the school district showed a low mean because it did a poor job of instruction in one field or another. This charge cannot be made, obviously, when the test is one that measures the general reaction of the children in the district to solving problems relating to his total environment and not the result of specific in-school, curriculum oriented knowledges and skills such as are measured by the achievement battery.

After all, it makes little difference whether one says that the result on a test such as the Otis-Lennon is due to environmental influences or to heredity or to some unknown and probably unknowable mixture of the two. The high and positive correlation between the results on the Otis and the results obtained when achievement tests are administered to the same children seems to establish, within all reasonable bounds, that there are factors independent of the school instructional program which gravely affect the amount of learning that takes place. Any comprehensive plan for statewide development in the curriculum to improve the quality of learning must surely take into account the district variability thus noted, without at the same time neglecting the also clearly demonstrable fact that even within the poorest of these communities there exists a wide variation in performance including levels of talent which would challenge the best teacher. In other words, the variability of pupil scores within the community is still very large when translated into standard scores even when the community mean is low or high.

The inclusion of cumulative percents in these tables makes possible other interesting analyses. Assuming that one has a community profile available, it is possible to interpret the status of the community as reflected in transformed mean scores into statements such as:

"Community X has a standard score mean of 47 on Otis-Lennon, which is comparable to the 23rd percentile for the 137 districts. In other words 23% of the 137 school districts shown have an

1/ Note: The computer program used to produce these graphs was developed by Mr. Donald Bailey of the University of New Hampshire, Bureau of Educational Research and Testing Services. It is an intermediate printout in a comprehensive procedure for obtaining local stanines and correlations.

Table III-D-1
Frequency and Cumulative Percent Distributions of Linear Standard Scores Corresponding to School District Means for 137 School Districts in New Hampshire Testing in the Fall of 1969

Grade 4

Otis-Lennon Mental Ability Test: Elementary II Battery: Form J

SCORE	PERCENTILE	STANINE	FREQUENCY	
31	1	1	1	I*
32	1	1	0	I
33	1	1	0	I
34	1	1	0	I
35	1	1	0	I
36	1	1	1	I*
37	1	1	0	I
38	2	1	1	I*
39	3	1	1	I*
40	4	1	1	I*
41	4	1	0	I
42	5	2	2	I***
43	7	2	3	I*****
44	9	2	3	I*****
45	16	3	9	I*****************
46	23	3	10	I********************
47	29	4	8	I****************
48	35	4	8	I****************
49	43	4	11	I**********************
50	69	5	26	I**
51	72	6	13	I*************************
52	83	7	16	I*******************************
53	90	7	9	I*****************
54	92	8	3	I*****
55	96	8	4	I*******
56	98	8	2	I***
57	98	9	2	I***
58	99	9	1	I*
59	99	9	0	I
60	99	9	1	I*
61	99	9	1	I*

Mean = 49.28

Standard Deviation = 4.28

Stanford Achievement Test: Intermediate I Battery: Form X
Word Meaning

SCORE	PERCENTILE	STANINE	FREQUENCY	
38	1	1	1	I**
39	1	1	1	I**
40	1	1	0	I
41	1	1	0	I
42	5	1	5	I**********
43	6	2	1	I**
44	9	2	5	I**********
45	14	3	6	I************
46	22	3	11	I**********************
47	28	4	9	I******************
48	35	4	9	I******************
49	45	4	13	I************************* *****
50	51	5	9	I******************
51	69	5	24	I**
52	77	6	11	I**********************
53	86	7	13	I**************************
54	93	8	10	I********************
55	97	8	4	I********
56	98	9	0	I
57	99	9	3	I******
58	99	9	0	I
59	99	9	0	I
60	99	9	1	I**
61	99	9	0	I
62	99	9	0	I
63	99	9	0	I
64	99	9	0	I
65	99	9	0	I
66	99	9	1	I**

Mean = 49.71

Standard Deviation = 4.03

Table III-D-1
(Continued)

Grade 4

Stanford Achievement Test: Intermediate I Battery: Form X

Paragraph Meaning

SCORE	PERCENTILE	STANINE	FREQUENCY	
37	1	1	1	Mean = 49.50
38	1	1	0	
39	1	1	0	
40	1	1	0	Standard
41	1	1	1	Deviation = 3.62
42	1	1	0	
43	4	1	4	
44	7	2	3	
45	11	2	4	
46	17	3	8	
47	27	3	14	
48	37	4	17	
49	51	5	18	
50	64	6	15	
51	74	6	13	
52	84	7	15	
53	89	7	7	
54	94	8	5	
55	96	9	2	
56	98	9	1	
57	98	9	4	
58	99	9	2	
59	99	9	0	
60	99	9	0	
61	99	9	0	
62	99	9	1	

Arithmetic Computation

SCORE	PERCENTILE	STANINE	FREQUENCY	
38	1	1	2	Mean = 48.82
39	1	1	0	
40	3	1	2	
41	6	2	4	Standard
42	7	2	2	Deviation = 4.51
43	11	2	5	
44	18	3	10	
45	24	3	8	
46	29	4	6	
47	37	4	11	
48	49	5	17	
49	59	5	12	
50	65	6	11	
51	72	6	8	
52	78	6	9	
53	87	7	12	
54	91	7	5	
55	94	8	5	
56	96	8	3	
57	98	9	2	
58	98	9	0	
59	98	9	1	
60	99	9	0	
61	99	9	0	
62	99	9	2	

Table III-D-1
(Continued)

Grade 4

Stanford Achievement Test: Intermediate I Battery: Form X
Arithmetic Concepts

SCORE	PERCENTILE	STANINE	FREQUENCY	
35	1	1	1	I**
36	1	1	0	I
37	1	1	0	I
38	1	1	0	I
39	1	1	1	I**
40	4	1	3	I******
41	6	2	3	I******
42	8	2	3	I******
43	12	2	5	I**********
44	15	3	5	I**********
45	20	3	6	I************
46	25	3	7	I**************
47	35	4	15	I******************************
48	42	4	8	I****************
49	52	5	14	I****************************
50	69	6	73	I**********
51	76	6	10	I********************
52	86	7	14	I****************************
53	91	7	6	I************
54	96	8	8	I****************
55	99	9	3	I******
56	99	9	0	I
57	99	9	0	I
58	99	9	0	I
59	99	9	1	I**
60	99	9	1	I**

Mean = 48.66

Standard Deviation = 4.03

Arithmetic Applications

SCORE	PERCENTILE	STANINE	FREQUENCY	
34	1	1	1	I*
35	1	1	0	I
36	1	1	0	I
37	1	1	0	I
38	1	1	0	I
39	1	1	0	I
40	1	1	1	I*
41	1	1	0	I
42	3	1	2	I**
43	7	2	5	I********
44	11	2	4	I********
45	15	3	6	I**********
46	20	3	7	I************
47	29	4	12	I**********************
48	37	4	11	I********************
49	50	5	17	I********************************
50	74	6	31	I**********
51	83	6	10	I******************
52	93	7	10	I******************
53	89	7	3	I****
54	93	8	5	I********
55	97	8	6	I**********
56	98	9	1	I*
57	98	9	0	I
58	99	9	1	I*
59	99	9	2	I***

Mean = 49.09

Standard Deviation = 3.67

Table III-D-2
Frequency and Cumulative Percent Distributions of Linear Standard Scores Corresponding to School District Means for 137 School Districts in New Hampshire Testing in the Fall of 1969

Grade 6

Otis-Lennon Mental Ability Test: Elementary II Battery: Form K

SCORE	PERCENTILE	STANINE	FREQUENCY	
24	1	1	1	I**
25	1	1	0	I
26	1	1	0	I
27	1	1	0	I
28	1	1	0	I
29	1	1	0	I
30	1	1	1	I**
31	1	1	0	I
32	1	1	0	I
33	1	1	0	I
34	1	1	0	I
35	1	1	0	I
36	1	1	0	I
37	1	1	0	I
38	1	1	0	I
39	1	1	0	I
40	1	1	0	I
41	2	1	1	I**
42	2	1	0	I
43	6	2	5	I**********
44	7	2	2	I****
45	10	2	4	I*********
46	15	3	7	I****************
47	23	3	11	I************************
48	34	4	15	I**********************************
49	42	4	11	I************************
50	58	5	21	I**
51	72	6	19	I**
52	80	6	12	I**************************
53	91	7	14	I*******************************
54	93	8	4	I*********
55	96	8	3	I*******
56	98	9	3	I*******
57	99	9	3	I*******

Mean = 49.50

Standard Deviation = 4.21

Stanford Achievement Test: Intermediate II Battery: Form X
Word Meaning

SCORE	PERCENTILE	STANINE	FREQUENCY	
39	1	1	1	I**
40	1	1	0	I
41	1	1	1	I**
42	1	1	1	I**
43	3	1	4	I**********
44	8	2	4	I**********
45	11	2	4	I**********
46	18	3	8	I********************
47	31	4	20	I**
48	40	4	12	I******************************
49	54	5	19	I**
50	64	5	13	I********************************
51	73	6	13	I********************************
52	83	7	14	I**********************************
53	90	7	9	I**********************
54	93	8	5	I************
55	96	8	3	I*******
56	97	9	2	I*****
57	99	9	2	I*****
58	99	9	2	I*****

Mean = 49.34

Standard Deviation = 3.46

Table III-D-2
(Continued)

Grade 6

Stanford Achievement Test: Intermediate II Battery: Form X
Paragraph Meaning

SCORE	PERCENTILE	STANINE	FREQUENCY	
40	1	1	1	I**
41	1	1	0	I
42	1	1	1	I**
43	2	1	1	I**
44	4	1	3	I*******
45	11	2	9	I**********
46	17	3	8	I*********
47	26	3	12	I**************
48	39	4	19	I**********
49	52	5	17	I**********
50	70	6	25	I**********
51	79	6	12	I**********
52	89	7	14	I**********
53	91	8	3	I*******
54	94	8	4	I*********
55	96	8	3	I*******
56	99	9	4	I*********
57	99	9	0	I
58	99	9	0	I
59	99	9	1	I**

Mean = 49.28

Standard Deviation = 3.05

Arithmetic Computation

SCORE	PERCENTILE	STANINE	FREQUENCY	
42	4	1	6	I**********
43	10	2	8	I*********
44	16	3	8	I*********
45	26	3	13	I**************
46	32	4	9	I**********
47	39	4	9	I**********
48	49	5	14	I***************
49	56	5	10	I**********
50	64	6	11	I**********
51	72	6	10	I**********
52	80	6	12	I**********
53	85	7	7	I*********
54	88	7	3	I*******
55	93	8	8	I*********
56	95	8	2	I*******
57	97	8	3	I*******
58	98	9	1	I***
59	98	9	0	I
60	99	9	1	I***
61	99	9	0	I
62	99	9	0	I
63	99	9	0	I
64	99	9	2	I*******

Mean = 49.04

Standard Deviation = 4.45

Table III-D-2
(Continued)

Grade 6

Stanford Achievement Test: Intermediate II Battery: Form X
Arithmetic Concepts

SCORE	PERCENTILE	STANINE	FREQUENCY	
39	1	1	1	I**
40	1	1	0	I
41	1	1	0	I
42	2	1	2	I*****
43	4	1	2	I*****
44	8	2	6	I**********
45	14	3	8	I**********
46	23	3	12	I**********
47	28	4	8	I**********
48	42	4	19	I**********
49	56	5	19	I**********
50	67	6	15	I**********
51	74	6	9	I**********
52	83	7	13	I**********
53	89	7	8	I**********
54	93	8	5	I**********
55	96	8	4	I**********
56	98	9	3	I*******
57	98	9	0	I
58	98	9	0	I
59	98	9	0	I
60	99	9	1	I**
61	99	9	2	I*****

Mean = 49.29

Standard
Deviation = 3.65

Arithmetic Applications

SCORE	PERCENTILE	STANINE	FREQUENCY	
33	1	1	1	I**
34	1	1	0	I
35	1	1	0	I
36	1	1	0	I
37	1	1	0	I
38	1	1	0	I
39	1	1	0	I
40	1	1	1	I**
41	1	1	0	I
42	2	1	1	I**
43	4	1	3	I*******
44	7	2	3	I*******
45	11	2	6	I**********
46	18	3	10	I**********
47	27	3	12	I**********
48	40	4	18	I**********
49	51	5	15	I**********
50	66	5	20	I**********
51	74	6	12	I**********
52	81	7	9	I**********
53	91	7	13	I**********
54	96	8	7	I**********
55	97	9	2	I*****
56	99	9	2	I*****
57	99	9	0	I
58	99	9	2	I*****

Mean = 49.30

Standard
Deviation = 3.51

Table III-D-3
Intercorrelations of District Means on Selected
Stanford Achievement Tests in Standard Score Form
Fall 1969

Grade 4

	Word Meaning	Para. Meaning	Arithmetic Computation	Arithmetic Concepts	Arithmetic Applications	Otis-Lennon
Word Meaning	1.00					
Para. Meaning	.84	1.00				
Computation	.54	.50	1.00			
Concepts	.69	.71	.67	1.00		
Applications	.66	.69	.67	.80	1.00	
Otis-Lennon	.79	.79	.51	.76	.73	1.00

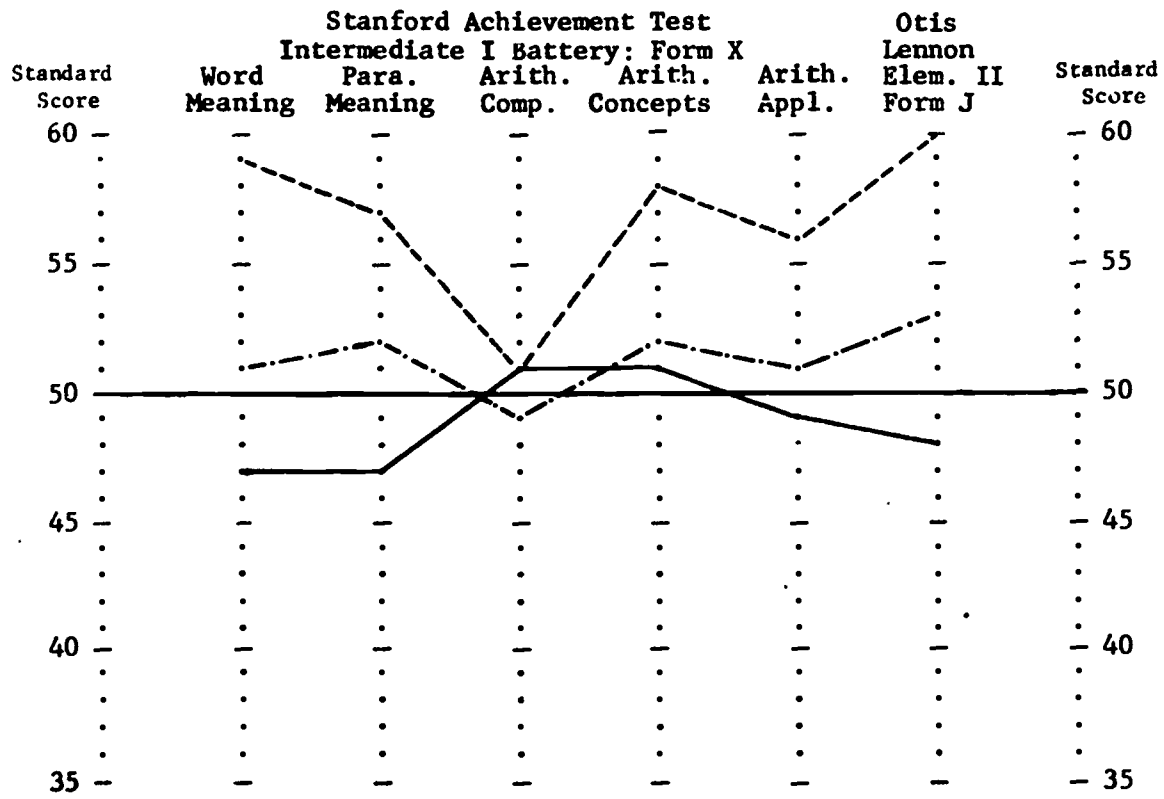
Grade 6

	Word Meaning	Para. Meaning	Arithmetic Computation	Arithmetic Concepts	Arithmetic Applications	Otis-Lennon
Word Meaning	1.00					
Para. Meaning	.84	1.00				
Computation	.37	.38	1.00			
Concepts	.66	.61	.65	1.00		
Applications	.62	.60	.63	.80	1.00	
Otis-Lennon	.62	.63	.34	.54	.52	1.00

Figure III-D-1

A SAMPLE PROFILE OF NEW HAMPSHIRE SCHOOL DISTRICT MEANS
Fall 1969

Grade 4



equal or lower standard score mean in measured mental ability."

In Table III-D-3 the intercorrelations between community means expressed as standard scores are given. As in all such correlation plots the 1.00 entries in the top left, lower right, first diagonal simply reflect the fact that the relationship between identical scores on a particular test would, of course, be 1, or perfect.

These intercorrelation tables can effectively be compared with the similar intercorrelations of the tests involved when pupil scores are used, not district means. These correlation coefficients tend to run somewhat lower because of a greater restriction in range. Perhaps the most significant line of figures on these tables is to be found in the bottom row where the correlations are reported between Otis-Lennon raw scores and each of the Stanford subtests included in this study.

Generally these correlations will be somewhat lower than the pupil correlations between the same pairs of tests, but the interesting fact is that these correlations are high as compared to many others reported in the table.

Attention is also called to the correlations between Arithmetic Concepts and the various other tests in the battery which also tend to run high, reflecting the fact that the Concepts test on Stanford is really a kind of indirect measure of mental ability. This is largely due to the fact that it reflects the knowledge of general principles taught as a part of the Mathematics program which are found to be very difficult by the less able pupils who have not reached a stage of maturity sufficient to permit them to work in terms of general principles.

One must recall that these data actually are correlations between means so that we are again emphatically reminded that a community or district has a character of its own not unlike the individual characteristics of a child and without any doubt this character of the community and of the schools within the community puts its stamp on the quality and effectiveness of the educational program. To put this differently and in more pragmatic terms, it just is not reasonable to expect a community which tends to run substantially below the population as a whole, which in this case is the State of New Hampshire, is going to achieve results on a standardized test up to the norm on the test. Another important concommitant, since we can infer a fairly close relationship between these test data and certain other sociological factors, is that larger proportions of Title I children will be found in the communities where the school district averages tend to be on the low side.

To emphasize how communities can differ without stressing the point unduly, a profile has been prepared using the linear standard scores on which are graphed the results for three communities selected by the writer as being representative of communities at the top, middle and lower parts of the distribution of standard scores corresponding to school district means. It would be possible to dwell on these profiles at some length, but for the purposes of this report this probably is not necessary. It is interesting, however, to point out that the community which is highest in terms of its average Otis-Lennon standard score is only doing average work in Arithmetic Computation. The writer leaves to the reader the task of making his own interpretation of the significance of this fact. (See Figure III-D-1 on page 32.)

SECTION IV

Description of the Title I Population From the "IN" and "OUT" Cards

The "IN" and "OUT" Cards as Control Documents

In Section II, delineating the changes made in the 1969-70 program, there is a discussion of the design and use of the "IN" and "OUT" cards.

One of the major advantages of the "IN" and "OUT" card procedure was the categorizing of Title I projects submitted and approved by the local communities. The following series of tables, from IV-1 to IV-8, summarizes the data for the Title I population obtained from the "IN" and "OUT" cards.

Table IV-1

In Table IV-1, the distribution of cases by category is shown. This table includes all Title I cases for whom "IN" cards were available. It is evident from the table that the Title I program is concentrated in the lower grades. (Data for Grades 3, 5, 7 and 9 are missing from this report since these grades were not involved in the testing program. The data are available, however.)

The data for Grades 2 and 4 indicate in excess of 800 cases in each instance, whereas the number of enrollees in the program in Grades 6 and 8 is between 400 and 500 cases, a substantial drop.

Most notable, also, is that the large majority of Title I children are involved in corrective and supportive reading programs. Even though the numbers of cases vary somewhat from grade to grade, the percentage of cases in corrective reading programs remains relatively constant, varying from a low of 64% in Grade 6 to a high of 81% in Grade 8. The percentage of the children in other projects listed according to our set of categories is so small that separate analysis is not justified. Hence, the only breakdowns used in this report will be for the total Title I group and, to the extent that it is possible, separate data analysis for those in the reading programs.

part at the beginning of the school year, with only a scattering of children coming into the program throughout the year. Thus it is apparent that the decision as to who shall be included in the program must have been made prior to the close of school in the previous year. This is also meritorious, assuming, of course, that the selection of these children has been made on the basis of adequate information.

Table IV -6

The duration of the Title I experience within the school year 1969-70 (Table IV -6) as indicated on the "OUT" card, varies greatly, although a majority of the children do stay in the program throughout the school year. Apparently some children are discharged from the program at various times, hopefully as their progress indicates that they are ready to move back into the regular stream of instruction.

This is in itself a good idea, provided, of course, that there is objective evidence at the time the child is removed from the program that he has indeed satisfied the goals set up for him at the beginning of the year.

The disadvantage of this is that if Fall and Spring testing is undertaken a substantial number of excused children, i.e., cases exiting before testing time, would not be tested in the Spring unless the school districts cooperate in carrying out the instruction that all Title I children, whether or not they had been discharged from the program, should be so tested. There is evidence that this was not done.

Table IV -7

In Table IV -7 we have an analysis of the reasons why pupils were discharged from the Title I project. The most obvious reason is, of course, that the school year was terminated and the inclusion or exclusion of a student from the program during the subsequent 1970-71 year had not yet been determined. The number of pupils leaving school for one reason or another, including transfers, appears to be negligible, but the percentages of those discharged from the Title I project because of satisfactory progress is fairly substantial. Using again the N for the tested group, the percentage of cases in Grade 4 is 7% and in Grade 6, 10%. In a sense, the number of individuals who are indicated as having made satisfactory progress or, better yet, the percent of these individuals as compared to the total group, is a measure of the success of the program if it is really directed toward remediation in some area such as reading or mathematics. The percent discharged is not impressive from this point of view, especially in reading. Pupils in the 4th and 6th grade, if the selection has been done carefully and the diagnosis is extensive, should respond to special help to the point where 50% or better would be able to be discharged from the program and sent back to their regular classes. This report does not utilize the U.S. Office of Education standard of "one year's progress in school for one year of school attendance" as a criterion because it is considered by this author to be totally unreasonable. However, the experience of this writer in directing a corrective reading program for Pinellas County, Florida, where careful selection followed by detailed analysis was made, indicates that the percentage could be as high as 75 or 80% under proper conditions of selection and analysis, a carefully prescribed program of remediation, and effective corrective teaching.

Table IV -8

In Table IV -8 , the teachers had an opportunity to indicate the level of progress made by each pupil in a Title I project. The table gives the number and percent of responses to each choice separately for the reading group and the total tested group. About 40 to 45% of both 4th grade pupils and 6th grade pupils were indicated as having made excellent progress, while another 40%+ were rated as having made modest but presumably significant progress. The percentage for whom only minor changes were evident or for whom no real benefit was reaped by the program is consistently small, not exceeding 14% of the total group in reading in Grade 4.

One must interpret these data as indicating a very optimistic outlook on the part of the teachers responsible for Title I instruction in light of the data comparing Title I pupils with the random sample tested Fall and Spring. Is it possible that expectation of progress is too low, so a small gain is credited too highly? One must draw his own conclusions after examining the data in this section.

Table IV -1

New Hampshire Statewide Testing Program 1969-70
Number and Percent of Cases Enrolled
in Each Project Category
Title I Program

Type of Project	GRADE 2		4		6		8	
	No.	%	No.	%	No.	%	No.	%
1. Language	31	3	17	2	11	2	16	3
2. Reading	657	68	606	74	286	64	398	81
3. Speech	69	7	26	3	18	4	4	1
4. Math	6	1	55	7	42	10	2	0
5. Guidance	79	8	56	7	27	6	19	4
6. Special Education	3	0	2	0	3	1	8	2
7. Psychiatric Services	14	2	0	0	0	0	0	0
8. Aides	50	5	27	3	22	5	6	1
9. Cultural Enrichment	16	2	25	3	5	1	25	5
10. Other	35	4	10	1	31	7	16	3
Totals	960	100	824	100	445	100	494	100

Table IV-2

New Hampshire Statewide Testing Program 1969-70
Distribution of Total Number of Hours of
Instruction for 1969-70 Title I Pupils
Separately for Reading and All Projects Combined

Grade 4

Hours/Week	Reading	All Cases
1	26	74
2	104	116
3	30	37
4	69	70
5	107	117
6	0	0
7	0	0
8	0	0
9	8	8
Full Time	0	6
Total No. of Cases	344	428

Grade 6

Hours/Week	Reading	All Cases
1	13	43
2	83	86
3	31	32
4	28	28
5	18	31
6	3	4
7	0	0
8	0	1
9	0	0
Full Time	0	13
Total No. of Cases	176	238

Table IV -3

Number and Percent of Pupils
By Type of Instructional Personnel Involved
Title I - 1969-70

Grade 4

Instructor	Reading No.	Reading %	Total Group No.	Total Group %
Reg. Classroom Teacher Only	0	0	0	0
Outside Person or Agency	0	0	7	2
Special Teacher in:				
Language	0	0	6	1
Reading	337	98	343	80
Speech	0	0	17	4
Math	0	0	10	2
Guidance	0	0	25	6
Aide	7	2	14	3
Other	0	0	6	1
Total	344		428	

Grade 6

Instructor	Reading No.	Reading %	Total Group No.	Total Group %
Reg. Classroom Teacher Only	0	0	13	6
Outside Person or Agency	0	0	0	0
Special Teacher in:				
Language	0	0	3	1
Reading	166	94	170	72
Speech	0	0	9	4
Math	0	0	8	3
Guidance	0	0	20	9
Aide	10	6	11	5
Other	0	0	1	0
Total	176		235	

Table IV-4

Number and Percent of 1969-70 Title I Pupils
Who Were in Title I Projects
In 1968-69 School Year

Grade 4

	Boy No.	Boy %	Girl No.	Girl %	Total* No.	Total* %
Reading						
Yes	94	27	41	12	135	39
No	102	30	95	28	198	58
Don't Know	8	2	1	0	9	3
	204	59	137	40	342	100
All Cases						
Yes	118	28	54	13	172	40
No	120	28	107	25	228	54
Don't Know	20	5	6	1	26	6
	258	61	167	39	425	100

Grade 6

	Boy No.	Boy %	Girl No.	Girl %	Total No.	Total %
Reading						
Yes	57	33	39	23	96	56
No	48	28	22	13	70	41
Don't Know	6	3	0	0	6	3
	111	64	61	36	172	100
All Cases						
Yes	83	36	52	22	135	58
No	53	23	31	13	84	36
Don't Know	9	4	5	2	14	6
	145	63	88	37	233	100

*Total Includes Pupil Who Did Not Code Sex

Table IV -5

Entry Date for Children
in the 1969-70 Title I Program (Tested Sample)

Grade 4

Entry Month	Reading			All Cases		
	Boy	Girl	Total	Boy	Girl	Total
September	137	79	216	162	92	254
October	15	8	23	23	12	35
November	0	2	2	0	2	2
December	13	11	24	13	12	25
January	2	3	5	2	3	5
February	2	0	2	2	0	2
March	8	10	18	8	10	18
April	1	0	1	1	0	1
May	0	0	0	0	0	0
June	2	5	7	2	5	7
Total	180	118	298	213	136	349

Grade 6

Entry Month	Reading			All Cases		
	Boy	Girl	Total	Boy	Girl	Total
September	75	45	120	88	54	142
October	2	4	6	4	8	12
November	6	0	6	6	1	7
December	9	3	12	10	3	13
January	3	2	5	3	2	5
February	1	0	1	1	0	1
March	1	1	2	1	1	2
April	0	0	0	0	0	0
May	0	0	0	0	0	0
June	0	0	0	0	0	0
Total	97	55	152	113	69	182

Table IV-6
Duration of Title I Experience
For All Available Cases Tested
in the Spring of 1970

Grade 4

Duration (Mos.)	Reading			All Cases		
	Boy	Girl	Total	Boy	Girl	Total
0 (No Out Cards)	3	5	8	3	5	8
1	0	0	0	0	0	0
2	10	2	12	10	2	12
3	19	19	38	20	19	39
4	12	7	19	12	7	19
5	3	4	7	3	5	8
6	4	2	6	7	3	10
7	6	2	8	11	2	13
8	66	48	114	69	51	120
9	57	29	86	78	42	120
Total	180	118	298	213	136	349

Grade 6

Duration (Mos.)	Reading			All Cases		
	Boy	Girl	Total	Boy	Girl	Total
0 (No Out Cards)	0	0	0	4	4	8
1	0	0	0	0	1	1
2	11	7	18	12	7	19
3	11	5	16	11	6	17
4	6	4	10	6	4	10
5	2	1	3	2	1	3
6	2	2	4	3	2	5
7	6	1	7	6	1	7
8	4	1	5	5	4	9
9	55	34	89	64	39	103
Total	97	55	152	113	69	182

Table IV-7

Reason for Termination of Participation
in 1969-70 Title I Project
Separately for Reading and Total Group

Grade 4

Reason	READING Boy No.	%	Girl No.	%	Total* No.	%
Satisfac.Progress	13	5	10	3	23	8
Left School	0	0	1	0	1	0
End of Sch. Year	163	55	102	35	256	90
Other	4	1	3	1	7	2
Total	180	61%	116	39%		100%

	TOTAL GROUP Boy No.	%	Girl No.	%	Total* No.	%
Satisfac.Progress	15	4	11	3	26	7
Left School	0	0	2	1	2	1
End of Sch. Year	190	55	118	34	309	89
Other	6	2	5	1	11	3
Total	211	61%	134	39%	348	100%

Grade 6

Reason	READING Boy No.	%	Girl No.	%	Total No.	%
Satisfac.Progress	8	5	7	5	15	10
Left School	0	0	0	0	0	0
End of Sch. Year	87	58	46	30	133	88
Other	1	1	2	1	3	2
Total	96	64%	55	36%	151	100%

	TOTAL GROUP Boy No.	%	Girl No.	%	Total No.	%
Satisfac.Progress	9	5	9	5	18	10
Left School	0	0	0	0	0	0
End of Sch. Year	102	56	58	32	160	88
Other	1	1	2	1	3	2
Total	112	62%	69	38%	181	100%

*Totals Include Pupils Who Did Not Code Sex

Table IV-8

Teacher Judgment
Concerning the Success of the Program
Title I 1969-70 Separately by Sex and
Separately by Reading versus Total Group

Grade 4

Success	READING Boy No.	%	Girl No.	%	Total* No.	%
Excell.Progress	79	27	53	17	133	44
Modest Progress	72	24	53	18	125	42
Minor Change	22	7	9	3	31	10
No Real Benefit	8	3	2	1	10	4
Total	181	61%	117	39%	299	100%

	TOTAL GROUP Boy No.	%	Girl No.	%	Total* No.	%
Excell.Progress	94	27	62	18	157	45
Modest Progress	87	25	62	18	149	43
Minor Change	24	7	10	3	34	10
No Real Benefit	8	2	2	0	10	2
Total	213	61%	136	39%	350	100%

Grade 6

Success	READING Boy No.	%	Girl No.	%	Total No.	%
Excell.Progress	39	26	26	17	65	44
Modest Progress	39	26	27	18	66	44
Minor Change	13	9	1	1	14	9
No Real Benefit	3	2	1	1	4	3
Total	94	63%	55	37%	149	100%

	TOTAL GROUP Boy No.	%	Girl No.	%	Total No.	%
Excell.Progress	47	26	36	20	83	47
Modest Progress	47	26	30	17	77	43
Minor Change	13	7	2	1	15	8
No Real Benefit	3	2	1	1	4	2
Total	110	61%	69	39%	179	100%

*Totals Include Pupils Who Did Not Code Sex

SECTION V

The Random Sample

The NEED for Random Sample Testing Program

The Stanford authors have provided no really meaningful way of comparing results in a test-retest situation over the relatively short period of time from Fall to Spring.1/ How then can one interpret such re-test data for special subgroups such as Title I? To provide for this situation, the New Hampshire 1969-70 statewide program under Title I auspices initiated the Spring re-testing of a random sample of cases selected from the entire population tested in the Fall to provide a state Spring "norm" group at each grade level.

The random sample was carefully drawn, using appropriate computerized statistical methods. Approximately 1500 children per grade from those tested in the Fall were identified to be retested.

Tests were provided by Title I for all of these children, but the number of cases actually tested or, at least for whom results were finally available for analysis, was typically less than 50% of those selected at each grade level.2/ This raised a serious question as to the representativeness of the TESTED random sample. However, when the paired cases from the random sample that were tested in the spring were finally available for both fall and spring and were compared with the total sample for fall testing, it was concluded that the differences, although some did exist, were not of practical significance and that the data for the partial random sample (now necessarily thought of as representative rather than random) provided a good guide as to the amount of gain to be expected by a cross-section group within this State over the seven month period between Fall and Spring testing, i.e., from October to May. Thus a much more realistic basis for evaluating Title I performance was made available than would otherwise have been the case.

Determining the Representativeness of the Tested Random Sample

The procedure for determining the representativeness of the tested random sample for any test included making a distribution of the scores for the selected cases from the Fall test results and plotting this Fall sample distribution on an Otis Normal Percentile Chart on which the total state Fall distribution was also plotted.

1/ They are not unique in this respect. Tests such as Stanford, California Achievement or Metropolitan etc. never were intended for retesting over short periods of time. A new technology is involved and test makers are hard at work on this problem.

2/ Some cases were lost who supposedly were tested because insufficient matching ID information was available.

It would not be sufficient to make this comparison on the basis of means and standard deviations alone or of any other simple set of statistics that did not describe the entire distribution. The only satisfactory way of making this comparison in a manner that would be clearly understood by anyone reading this report was to compare the distributions graphically, at least for some of the tests involved. The Otis Normal Percentile Chart is normal probability paper prepared especially for plotting distributions of test scores where it is desired to determine first, whether the distribution approximates a normal curve, and secondly, to compare a number of distributions which represent samples presumably comparable.

This was done first for the Otis-Lennon Mental Ability Test deviation IQs (DIQ) and the results for Grade 4 and Grade 6 are shown on the Normal Percentile Charts labeled Charts V-1 and V-2. Unit increments of one point are plotted and the line was drawn by connecting plotted points, thus allowing the greatest possible variation from one graphed line to the other. Smoothed lines are often suspect unless some statistically fairly precise way of smoothing is used or the person who does the graphs has had extensive experience at this task. In this situation, such refined smoothing seemed superfluous.

In view of the relatively small number of cases in the tested random sample, one would naturally expect a less smooth curve for the sample than for the total populations of about 11,000 or 12,000 cases tested in the fall. The general trend of the plotted line for the tested random sample, however, did reveal some slight systematic differences between the total group and the small sample.

The "tails" of the graphed distributions for both grades reproduced in this report have been curtailed for reasons of space. The total range of IQs for the state is shown in tabular form in another part of this report.

Looking first at the plotted line for the total sample, it will be seen that this line approximates a straight line for the major part of its length. The tested random sample also is reasonably straight but shows a tendency to be somewhat above the total population, especially in the lower part of the curve, but approaches the state graph more closely at the top of the curve. Minor deviations must be considered to be of negligible importance for the purpose to be served here.

Our conclusion from a study of the Grade 4 chart is that the tested sample is slightly superior in measured "brightness" to the total population, especially in the lower range. At the median the difference amounts to about a point and a half, while at the 10th percentile this difference is perhaps two and a half points in terms of DIQ.

The small number of cases available in the sample did not permit any deletion of cases to force the distribution for the tested random sam-

ple in line with the total population tested, and therefore we tentatively concluded that the best procedure was to accept this sample as reasonably representative and study the amount of gain from Fall to Spring for this particular group as a basis for comparison of the gain made by the Title I children in this same grade, including all Title I cases regardless of the type of project in which they might be found and, if possible in terms of available resources, for those in reading projects separately.

In Chart V -2 for Grade 6 the graphed line for the total sample looks very similar to that for Grade 4, but the random sample ever more closely approximates the total group than in the case of Grade 4. Thus we can say that for both Grades 4 and 6 the measured brightness of the tested random sample certainly seems to be generally representative of what might be expected had the entire state been retested with Otis-Lennon in the spring. This is especially true because of the nature of the Otis-Lennon test and because of the conditions under which the sample was selected. The tested random sample was not even identified for Spring testing purposes until nearly time for the tests to be administered in May.

Graphs similar to those shown were prepared for the five major Stanford subjects on which we are concentrating our attention in making comparisons of Fall and Spring testing for both the random sample and Title I. All charts contained a line for the entire state and for the tested random (representative) sample, both Fall and Spring. These tests are as follows: Word Knowledge, Paragraph Meaning, Arithmetic Computation, Arithmetic Concepts, Arithmetic Applications. None of the achievement test graphs can be shown, for reasons of economy, but the selected statistics extracted from the graphs are presented elsewhere. In summary, all such charts supported our belief that the random (now representative) sample fairly reflected the state performance in each comparison made.

In conclusion, it might be helpful to approach this evaluation of the random sample from a slightly different point of view. The tested random (representative) sample constitutes a group of children for whom there is no known reason to suspect systematic bias such as Hawthorne effect. The growth achieved by these children in the tested random (representative) sample over the seven month period is certainly one realistic touchstone as to the amount of growth to be expected during such a period under conditions existing in New Hampshire. No better data exist as the basis for such comparisons. Statistical niceties may be lacking but common sense is not. If lack of cooperation, logistical support, and indifference prevented a more precise comparison group this is indeed unfortunate but not the fault of this writer or of the Title I staff. It does, however, highlight the need for a deeper understanding of evaluative research on the part of all those who wish to know, <u>really</u> wish to know, whether their efforts are availing.

In Table V-1 selected percentiles are shown, as read from the available Normal Percentile Charts, for the total state and for the tested random sample. No statistical significance tests have been made for these data and none are sensible under the circumstances. Values have been read to the nearest whole number since fractional values are of little use and suggest a precision not resident in the data.

Table V-2 shows the means and standard deviations for raw scores on selected achievement tests, discussed in this report, administered in the fall of 1969. Casual inspection is sufficient to establish rather clearly that the random sample is comparable to the total population on these parameters. This is an instance where tests of statistical significance might be applied if it were not for the fact that the random sample ceased to be random when substantial numbers of pupils failed to take the tests in the spring and, therefore, were excluded.

If one wanted to indulge in an hypothetical exercise one could consider the total population tested in the fall as the universe, which would mean that the means and standard deviations reported are free of sampling error. It then would be possible to test the extent to which the representative sample with which we are left might be significantly different if the assumption of randomness was true. Since the writer can see no valid purpose for doing this, such statistical tests of significance have not been carried out.

Neither Otis-Lennon nor composite prognostic scores have been included in this comparison because of failure of the service centers to provide the necessary information.

Otis-Lennon IQ | State-wide 11,92[illegible] | Random Sample 585

Chart V-1

NORMAL PERCENTILE CHART

By Arthur S. Otis

Score Intervals	Otis-Lennon IQ	State-wide	Random Sample	Percentile Ranks
131+	259	2.2	7	99
130	252	2.1	2	99
129	245	2.1	4	98
128	238	2.1	3	98
127	231	2.0	2	97
126	224	1.9	3	97
125	217	1.8	5	96
124	210	1.8	7	96
123	203	1.7	4	94
122	196	1.7	6	93
121	189	1.6	8	92
120	182	1.5	5	91
119	175	1.5	7	90
118	168	1.4	7	87
117	161	1.4	12	83
116	154	1.3	10	76
115	147	1.2	10	67
114	140	1.2	11	57
113	133	1.1	17	46
112	126	1.1	9	29
111	119	1.0	17	20
110	112	0.9	17	12
109	105	0.9	3	7
108	098	0.8	7	4
107	091	0.8	2	3
106	084	0.7	12	2
105	077	0.7	12	1
104	070	0.6	2	1
103	063	0.5	1	1
102	056	0.5	23	1
101	049	0.4	8	1
100	042	0.4	24	1
99	035	0.3	18	1
98	028	0.3	16	1
97	021	0.2	15	1
96	014	0.1	3	1
95	007	0.1	1	1
94	000	0.0	3	1
93	993	0.0	1	1
92	986	0.0	1	1
91	979	0.0	1	1
90	972	0.0	1	1
89	965	0.0	1	1
88	958	0.0	1	1
87	951	0.0	1	1
86	944	0.0	1	1
85	937	0.0	1	1
84	930	0.0	1	1
83	923	0.0	1	1
82	916	0.0	1	1
81	909	0.0	1	1
80	902	0.0	1	1
79	895	0.0	1	1
78	888	0.0	1	1
77	881	0.0	1	1
76	874	0.0	1	1
75	867	0.0	1	1
74	860	0.0	1	1
-73	853	0.0	1	1
Medians		101	102	
Standard Deviation Scale				

PERCENTILE SCALE

1 2 3 4 5 10 20 30 40 50 60 70 80 90 95 96 97 98 99 99.2 99.4 99.6 99.8

Graphic Representation of the Distribution of Otis-Lennon Mental Ability Test Deviation IQs for Total State and Tested Random Sample
Fall 1969 Grade 4

State———— Random Sample– – – – –

-3σ -2σ -1σ M +1σ +2σ +3σ

Otis-Lennon IQ | Statewide 11,549 | Random Sample 641

Chart V-2

NORMAL PERCENTILE CHART

By Arthur S. Otis

Score intervals	I Freq.	I Sub-total	I Per cent	II Freq.	II Sub-total	II Per cent
131+	1103		98	29		97
130	32		97	3		97
129	101		97	3		91
128	17		96	3		96
127	167		96	3		96
126	0		95	2		95
125	115		95	15		95
124	2		93	3		93
123	170		93	9		93
122	32		91	2		91
121	203		90			91
120	151		89	2		89
119	201		88	11		88
118	226		86	19		87
117	199		84	3		84
116	252		83	24		83
115	212		81	12		80
114	208		79	10		78
113	211		77	11		76
112	202		76	10		74
111	193		74	14		73
110	187		71	14		71
109	176		69	9		68
108	175		66	13		67
107	156		64	13		65
106	164		62	15		63
105	134		59	26		61
104	199		56	21		57
103	205		54	22		53
102	209		51	21		50
101	207		48	11		47
100	191		45	12		44
99	181		42	22		41
98	96		39	11		39
97	216		37	16		36
96	122		34	16		34
95	121		31	14		32
94	294		29	12		27
93	210		26	13		26
92	210		24	14		23
91	224		22	14		21
90	123		19	1		18
89	241		15	11		17
88	175		16	13		15
87	177		14	7		13
86	127		13	3		12
85	138		11	6		11
84	99		11	4		10
83	114		10	2		9
82	92		9	2		9
81	75		8	2		8
80	80		7	2		8
79	80		7	5		7
78	73		6	4		6
77	62		5	4		6
76	51		5	2		5
75	47		4	3		5
74	45		4	2		4
-73	414		41	26		4
Medians	102			102		

Graphic Representation of the Distribution of Otis-Lennon Mental Ability Test Deviation IQs for Total State and Tested Random Sample
Fall 1969 Grade 6

State _____ Random Sample - - - - -

PERCENTILE SCALE

1 2 3 4 5 10 20 30 40 50 60 70 80 90 95 96 97 98 99 99.2 99.5 99.8

Standard Deviation Scale -3σ -2σ -σ M +σ +2σ +3σ

Published by World Book Company, Yonkers-on-Hudson, New York, and Chicago, Illinois.

Table V-1

A Comparison in Terms of Raw Scores of the
State-wide Populations in Grades 4 and 6 Tested in the Fall of 1969
with Fall Results for the Random (Representative) Sample
Subsequently Selected for Spring Re-testing as a Control Group

		Grade 4			Grade 6		
Test	%ile Rank	Total State	Random Sample	Diff.	Total State	Random Sample	Diff.
Word Meaning	75	19	21	+2	31	32	+1
	50	14	15	+1	24	26	+2
	25	9	10	+1	18	19	+2
Paragraph Meaning	75	29	30	+1	40	40	0
	50	22	23	+1	31	33	+2
	25	16	17	+1	23	25	+2
Arithmetic Computation	75	14	14	0	17	17	0
	50	11	11	0	13	13	0
	25	8	8	0	9	10	+1
Arithmetic Concepts	75	16	16	0	16	16	0
	50	12	12	0	12	13	+1
	25	8	9	+1	9	9	0
Arithmetic Applic.	75	16	16	0	22	22	0
	50	12	12	0	16	17	+1
	25	8	9	+1	12	12	0

Table V-2
A Comparison of Means and Standard Deviations
For the Random Sample and the Total Population
Fall 1969

Grade 4

		RAW SCORE MEAN		STANDARD DEVIATION	
SAT:Int.I:X	No.of Items	Random Sample	Total Population	Random Sample	Total Population
Word Meaning	38	15.9	15.0	7.1	7.0
Para. Meaning	60	24.4	23.4	9.4	9.4
Arith. Comp.	39	11.5	11.6	4.5	4.6
Arith. Concepts	32	12.9	12.7	5.2	5.3
Arith. Appl.	33	12.9	12.7	5.1	5.3

Grade 6

		RAW SCORE MEAN		STANDARD DEVIATION	
SAT:Int.II:X	No.of Items	Random Sample	Total Population	Random Sample	Total Population
Word Meaning	48	25.7	25.0	8.5	8.7
Para. Meaning	64	32.9	32.2	11.2	11.5
Arith. Comp.	39	13.9	13.6	5.5	5.5
Arith. Concepts	32	13.5	13.3	5.1	5.3
Arith. Appl.	39	17.6	17.4	6.8	6.9

SECTION VI

The Tested Title I Population in New Hampshire Described and Compared with the Random (Representative) Sample

In this section we will present certain data describing the Title I population in Grades 4 and 6 in comparison with the random (representative) sample for these same grades. First, however, we must note an exception which renders all comparisons in this study somewhat moot as a description of the situation for Title I as a whole in this state. Notice the large discrepancy in cases within the Title I samples at Grade 4 and Grade 6. In Grade 4, 431 cases are included in comparison with 230 cases in Grade 6. These figures in neither case represent the entire Title I population in these grades but the population of Title I cases in Grade 6 for whom we have complete test data is a seriously biased population. Grade 4 is also biased but probably not as extensively since it probably represents a larger proportion of Title I cases in that grade. Certain significant and constant differences between the two grade populations seem to run through all the comparisons made subsequently. Certain communities, especially some of the larger cities, are not included in the study because they failed to test either fall or spring. These data simply are NOT representative of all New Hampshire Title I cases but, since they are a defined sample, the basic descriptive statistics relating to factors as significant as chronological age range, etc., probably reflect the degree of unrepresentativeness of the Title I population as regards the total group reasonably well.

The question must arise as to how this situation comes about. Recall that the basic law says that every school district is entitled to Title I funds in proportion to the number of families below the poverty level plus the number of families receiving aid-to-dependent-children, etc. It also requires that every project within Title I shall be evaluated, but it does not specify that this evaluation is to be dictated by the central office for Title I in the state, either as to the instruments used or as to the methods of analysis. Therefore, in this State, at least, (and probably in many others) any attempt to analyze data for Title I children obtained by voluntary cooperation, using prescribed instruments, is bound to be a biased description of what is going on.

In New Hampshire, some of the larger cities chose to do their own evaluations, using tests which they determined, using methods of analysis which they determined, and reporting only such data as they were directed to produce by the State Title I office. This consisted essentially of data describing the Title I population and did not include individual test results. There was no prescribed constraint that they must describe their Title I population with respect to their own total population. National norms were considered sufficient and assumed to be comparable from test to test. No effort has been made, to this writer's knowledge, to summarize all such information received from these non-conforming communities. In other words, no one knows how well these various communities carried out their Title I project obligations as measured by a uniformly prescribed test. Indeed, there may be some shining gems of masterly accomplishment in achieving great success with their Title I children which are not reflected in the data we are describing.

With these limitations clearly in mind, let us take a look at the data we have for the tested sample.

Chronological Age

Perhaps one of the basic parameters that should always be examined is the characteristics of the population with respect to chronological age related to grade in a graded type of school organization. In order to provide the necessary comparative information, in the remainder of this report only the random (representative) sample will be compared with the Title I population since we only have data for both Fall and Spring for these groups.

Note first the range of chronological ages for the random sample (Table VI-1 and Table VI-2, which is essentially identical with that for the entire group tested in the fall. The percent of boys versus girls varies slightly from Grade 4 to Grade 6. In Grade 4, 48% are boys and 52% are girls. In Grade 6, 53% are boys and only 47% are girls.

Compare now the similar information for the Title I children as tested. In Grade 4, 61% are boys and 39% are girls and in Grade 6, 62% are boys and 38% are girls. It is rather remarkable that the percentage of boys in Title I is so constant from Grade 4 to Grade 6, but it is even more remarkable that these figures check very closely with many kinds of evidence as to the percent of boys who have difficulty of one sort or another in school as compared to the percent of girls. Time will not be taken to document this fact in this study but any reader wishing to do so can find many bits of evidence to indicate that it is boys who generally have reading problems; it is boys who generally get in trouble with the law and are ruled to be juvenile delinquents; it is boys who generally tend to have emotional difficulties of various kinds requiring their referral to some outside agencies. Furthermore, the proportion of boys to girls is more or less the same as reported here for chronological age.

The median age of boys in the representative sample for Grade 4 is 9-6, for girls 9-5, with a median age of 9-6 for the entire sample. For Title I the median age is 9 years 11 months for boys, 9 years 8 months for girls and the median age for the total Title I sample is 9-10. Title I children in Grade 4 thus average four months older than the total population. For Grade 6 the comparable

Table VI-1

New Hampshire Statewide Testing Program - Fall 1969
Distribution of Chronological Ages
Separately by Sex and for the Total Group
Random Sample and Title I

Grade 4

Age in Years & Months	Random Sample			Title I		
	Boys	Girls	Total*	Boys	Girls	Total*
14- 4 to 14- 9	0	0	0	1	0	1
13-10 to 14- 3	0	0	0	0	0	0
13- 4 to 13- 9	1	0	1	1	0	1
12-10 to 13- 3	0	0	0	0	0	0
12- 4 to 12- 9	0	0	0	1	0	1
11-10 to 12- 3	1	2	3	1	4	5
11- 4 to 11- 9	3	3	6	12	1	13
10-10 to 11- 3	1	4	6	26	8	34
10- 9	2	1	3	11	4	15
10- 8	1	1	2	5	5	10
10- 7	1	3	4	7	2	9
10- 6	1	3	4	5	1	6
10- 5	6	6	12	12	6	18
10- 4	4	1	7	9	2	11
10- 3	3	2	5	12	5	17
10- 2	6	6	13	6	6	12
10- 1	7	7	15	8	8	16
10- 0	7	3	10	8	6	14
9-11	10	4	14	11	9	20
9-10	9	8	18	12	3	15
9- 9	19	26	45	7	9	16
9- 8	19	20	41	12	9	21
9- 7	25	21	48	12	13	25
9- 6	20	21	43	7	4	11
9- 5	18	19	37	7	6	13
9- 4	16	17	36	7	8	15
9- 3	14	18	34	13	10	23
9- 2	14	23	39	15	10	25
9- 1	15	22	39	12	9	21
9- 0	13	19	32	13	6	20
8-11	19	19	38	5	11	16
8-10	13	16	31	3	1	4
8- 4 to 8- 9	0	0	0	0	2	2
7-10 to 8- 3	0	0	0	0	1	1
Total Ns	268	295	586	261	169	431
N Mid-12 Months	205	241	463	128	98	225
N Mid-18 Months	247	271	538	171	134	313
% Mid-12 Months	76.5	81.7	79.0	49.0	58.0	52.2
% Mid-18 Months	92.2	91.9	91.8	65.5	79.3	72.6
Median Age	9-6	9-5	9-6	9-11	9-8	9-10

*Includes Students Who Did Not Code Sex

See Section III-A, p. 5 and the Metropolitan Manual fo Interpreting, p. 2, 9-11, and 23-24 for further discussion of the age-controlle sample.

Table VI-3
New Hampshire Statewide Testing Program
Distribution of Otis-Lennon Deviation IQs
Separately for Boys, Girls, and Total Group
Tested Fall 1969
RANDOM SAMPLE

	GRADE 4						GRADE 6					
	BOYS		GIRLS		TOTAL*		BOYS		GIRLS		TOTAL*	
IQ Interval	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age
150	1	99			1	99						
147-149	0	99			0	99						
144-146	0	99			0	99						
141-143	0	99			0	99	1	99	3	99	4	99
138-140	2	99			2	99	0	99	3	99	3	99
135-137	1	99			1	99	2	99	0	98	3	99
132-134	1	98	1	99	2	99	3	99	6	98	9	98
129-131	3	98	3	99	7	99	5	98	3	96	8	97
126-128	4	97	4	99	8	98	2	96	3	95	5	96
123-125	5	95	11	97	16	96	15	96	8	94	24	95
120-122	7	94	11	94	19	94	7	91	11	91	18	91
117-119	10	91	15	90	26	90	16	89	16	88	33	88
114-116	14	87	20	85	36	86	27	84	17	82	46	83
111-113	20	82	22	78	43	80	16	76	19	76	35	76
108-110	21	75	26	71	47	73	16	71	17	70	36	71
105-107	21	67	23	62	48	64	21	66	29	64	54	65
102-104	23	59	28	54	53	56	25	59	39	55	64	57
99-101	23	50	27	44	50	47	26	51	20	41	48	46
96-98	19	42	24	35	44	39	30	43	29	35	61	39
93-95	14	34	22	27	40	31	20	34	23	25	44	30
90-92	13	29	20	20	34	24	23	28	13	17	37	23
87-89	18	24	15	13	34	18	21	21	11	13	34	17
84-86	16	18	8	8	24	13	10	14	5	9	15	12
81-83	5	12	4	5	10	9	6	11	4	7	10	9
78-80	7	10	6	4	14	7	7	9	5	6	14	8
75-77	8	7	1	2	9	4	4	7	6	4	10	6
72-74	4	4	2	1	8	3	4	6	0	2	4	4
69-71	4	3	1	1	5	2	4	5	3	2	7	3
66-68	0	1	1	1	1	1	2	3	1	1	3	2
63-65	2	1			2	1	4	3	3	1	7	2
60-62	0	1			0	1	1	2			1	1
57-59	0	1			0	1	1	1			1	1
54-56	1	1			1	1	2	1			2	1
51-53							1	1			1	1
Totals	267		295		585*		322		297		641*	
Q3 %ile 75	110		112		111		112		113		113	
Q2 %ile 50	101		103		102		100		103		102	
Q1 %ile 25	89		94		92		91		95		93	
Mean	100.77		103.29		102		101.03		103.85		102.42	
Standard Dev.	14.93		12.39		13.76		15.76		14.35		15.11	

*The Distributions of DIQs for Boys and Girls do not sum to the Total Distribution because of the failure of a substantial number of pupils to code sex.

Table VI-4
New Hampshire Statewide Testing Program
Distribution of Otis-Lennon Deviation IQs
Separately for Boys, Girls, and Total Group
Tested Fall 1969
TITLE I

	GRADE 4						GRADE 6					
	BOYS		GIRLS		TOTAL*		BOYS		GIRLS		TOTAL*	
IQ Interval	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age	No.	Cum. %age
141-143									1	99	1	99
138-140									0	99	0	99
135-137			1	99	1	99			0	99	0	99
132-134	1	99	0	99	1	99			0	99	0	99
129-131	0	99	0	99	0	99			0	99	0	99
126-128	0	99	1	99	1	99	1	99	1	99	2	99
123-125	0	99	0	99	0	99	0	99	0	98	0	99
120-122	1	99	1	99	2	99	0	99	1	98	1	99
117-119	0	99	1	98	1	99	1	99	3	97	4	98
114-116	1	99	0	98	1	99	3	99	0	93	3	97
111-113	4	99	2	98	6	98	2	96	1	93	3	95
108-110	4	97	2	96	6	97	4	95	5	92	9	94
105-107	4	96	4	95	8	95	1	92	5	87	6	90
102-104	10	94	5	93	15	94	9	91	5	81	14	87
99-101	9	90	10	90	19	90	8	85	10	75	18	81
96-98	13	87	13	84	26	86	11	79	6	64	17	73
93-95	26	82	17	76	43	80	20	72	8	57	28	66
90-92	20	72	17	66	37	69	11	57	9	48	20	54
87-89	28	64	19	56	48	61	14	50	12	38	26	45
84-86	32	53	14	45	46	50	13	40	7	25	20	34
81-83	23	40	19	37	42	39	6	30	3	17	9	25
78-80	22	32	11	25	33	29	11	26	6	13	17	21
75-77	14	23	15	19	29	21	5	18	2	7	7	14
72-74	14	18	7	10	21	15	8	15	1	4	9	11
69-71	8	12	2	6	10	10	4	9	0	3	4	7
66-68	9	9	3	5	12	7	1	6	2	3	3	5
63-65	8	5	0	3	8	4	0	6	0	1	0	4
60-62	1	2	3	3	4	3	3	6	1	1	4	4
57-59	2	2	1	1	3	2	1	4			1	2
54-56	2	1	1	1	3	1	2	3			2	2
51-53	1	1			1	1	1	1			1	1
48-50							1	1			1	1
Totals	257		169		427*		141		89		230*	
Q3 %ile 75	93		95		94		97		101		98	
Q2 %ile 50	85		88		86		89		92		91	
Q1 %ile 25	78		80		79		80		86		83	
Mean	85.76		87.95		86.63		88.65		94.39		90.87	
Standard Dev.	12.37		12.26		12.35		13.69		13.16		13.75	

*More Title I pupils were tested in the fall than in the spring but only matched cases are included.

figures in the random sample are 11-7 for boys, 11-6 for girls and 11-6 for the total group, while the figures for Title I at Grade 6 are 11-11 for boys, 11-8 for girls, with a median age for the total Title I tested sample of 11 years and 10 months, again a difference of four months.

In these tables, the middle 18-month range has been set off from the rest of the distribution to represent children who are substantially at grade for age. This, in this writer's nomenclature is called the age controlled sample. This has been determined for the random sample as well as for the total fall distribution of chronological ages. In Grade 4 it includes children from 8 years and 10 months to 10 years and 3 months, a net range of one year and six months. In Grade 6 it includes children 10 years and 10 months to 12 years and 3 months. In other words, the ranges for Grades 4 and 6 differ from each other by exactly two years at the terminal points. In the representative sample about 92% of all of the cases fall within this age controlled range and only 8% of the representative sample are older than the uppermost bound of the 18-month range. In the Title I group, in Grade 4, 29% fall above the upper bound of the age controlled sample for the state as a whole and only three cases, all girls, are younger than the youngest child in the age controlled group. This means that the percent of retardation in the Title I population is very substantially larger than for the group as a whole, which is consistent, of course, with the finding that the median chronological age is substantially higher for Title I children in Grade 4 than for the group as a whole.

The statistics for Grade 6 are consistent. 12% of the children in the population are older than the upper bound of the age controlled sample, while in Title I 29% are older.

In conclusion then, we can describe the Title I population as generally being older; as having essentially the same spread of chronological ages as the total group but with a much larger proportion in the upper or older age brackets than is true of the state. Since it is obvious that children above the upper bound of the age controlled sample must have been retarded at least one year, we can say that the percent of children actually retarded in school for their grade placement is roughly 30% for both Grades 4 and 6.

The Random Representative Sample versus Title I Sample in Terms of Measured Mental Ability

In Tables VI-3 and VI-4 descriptive information is given concerning the measured mental ability of the random sample and Title I, using the Otis-Lennon Mental Ability Test deviation IQ (DIQ). Looking first at Table VI-3, which describes the random sample, we find that the distribution of DIQs in Grade 4 ranges from the 50's to about 150 and that the median IQ is 101 for boys, 103 for girls and 102 for the tested population. The same information for Grade 6 for the representative sample, shows medians of 100 for boys, 103 for girls, and 102 for a total. The distributions are essentially symmetrical and nearly normal and correspond very closely to comparable information for the total state presented elsewhere. Thus these data confirm our earlier conclusion that the New Hampshire population is at or slightly above the national norm sample on the Otis-Lennon.

Now looking at the data for Title I (Table VI-4) we see a substantial contrast. The median IQ for boys in the tested Title I group in Grade 4 is only 85, for girls 88, and for the total group 86. In Grade 6, the median is 89 for boys, 92 for girls, with 91 as the median DIQ for the total Title I group at this grade level.

All these tables provide cumulative percentages so it is possible to tell by consulting any table what percent of children fall above any given DIQ level. For example, the cumulative percents describing the Title I population in Grade 4 show that 90% of the youngsters in this group have deviation IQs of 101 or lower or conversely only about 10% exceed that near normal median value for the state. Recall that the typical value found for the entire State of New Hampshire was 102 at Grade 4. At Grade 6 the comparable figure is 81% having DIQ's of 101 or lower with 19% having DIQ's higher than this. In other words, only 19% of the Title I children tested in Grade 6 had higher DIQ's than the average level of brightness for the state as a whole.

When all of these data are thoughtfully examined one can reach the conclusion that the Title I group is definitely a selected group both with respect to chronological age, and also for DIQ. The Title I populations in both grades are definitely over-age, slow-learning groups. Looking again at the cumulative percentages, we see that 15% of the Grade 4 Title I children have DIQ's on the Otis-Lennon Test of 74 or lower, while at Grade 6, 11% fall in this category. By contrast, for the tested random (representative) sample, only 3% of the children in Grade 4 have DIQ's of 74 or lower, while at Grade 6 the percent is 4%.

These data provide us with an opportunity to ask a very interesting question. Is it the intent of the Title I law to provide special help for slow-learning children in contrast to those who have corrective remedial defects in basic skills areas such as reading and math? I am quite sure the intent of the law is not clearly one or the other but the net result of the method of selection in New Hampshire, at least, is the choice of a group of relatively slow-learning children who are over-age for their grade, for whom the main task would appear to be to provide content of instruction suitable to their level of mental development in a sequence and at a rate of presentation suitable to their somewhat slower learning pace. Most significantly this says by definition almost, that it is unreasonable to expect development in the basic skills areas at a rate commensurate with the normal rate, i.e., one year's growth for one year's life experience in school.

The data argue strongly for instruction oriented to the needs and learning ability rate of each individual child in any Title I project in this State. In the subsequent sections, comparative data will be presented concerning what actually took place by way of learning within the skills areas as measured by the Stanford Achievement Test over a period of seven calendar months roughly from October 15 to May 15.

To try to escape these conclusions by arguing that Otis-Lennon does not measure "learning potential" or "learning ability" is only to quibble. The test content is obviously not curriculum oriented especially at Grade 4. The scores on the test tend to correlate more highly with measures of achievement than any other test. Those identified by the test as slow learners are so identified by observation before testing. What the test does is to quantify this factor to permit relating it to other variables, - not perfectly perhaps but surprisingly well for its time limits and length in terms of number of items.

The argument put forward by the authors that the test measures "G" in the Spearman sense of the term is interesting but irrelevant to this discussion. At this point the writer couldn't care less how the output of the test is labelled. He cares very much that it does validly describe the pupils in the State of New Hampshire at these two grade levels in a manner consistent with his own 20+ years of experience with this population and that it describes the tested Title I sample in a logically consistent manner. 1/

1/ These data are remarkably consistent with Statewide 8th grade information collected by the writer using the precursor of the Otis-Lennon, namely the Otis Quick Scoring: Gamma, Test of Mental Ability in 1963 and 1964.

SECTION VII

Single-Variable Comparisons of Fall-Spring Performance for the Random Sample and for Title I Cases

Part A
Some Basic Measurement Problems

This is certainly not the place to enter into a lengthy discussion as to the nature of measurement in education and psychology. It is generally accepted that every measure of every kind, even those that appear on the surface to be quite precise, does include an error factor which is affected by many influences that are essentially unknown and unknowable. Such things as the quality of the test items in the sense of freedom from ambiguity, the length of the test, the applicability of the test to the local situation in terms of instructional validity, the quality of the test administration, the general emotional environment in which the tests are administered along a continuum that might go from stressful and emotionally upsetting to accepted and casual, - all these things and many others affect the performance of an individual on the day(s) he happens to take a test or a series of tests.

The trsnsformation of the raw score he earns, usually the number right, into a standard score does not in any way diminish or correct for these random error factors, although it may change the magnitude of the computed estimate of error because the standard deviation is arbitrarily altered.

Note that these errors of measurement (SE_m) are present even when a single test is given and only one score available. In fact the SE_m is primarily useful to give the user an idea of how much dependence he can put on a given test score. These errors vary in magnitude from one subtest to another within a battery depending upon the standard deviation of the raw scores and, particularly, the reliability of the test.

Standard errors of measurement in terms of raw scores are definitely not comparable from test to test.

When one test or test battery is given, let us say in the Fall, and this same test or an alternative form is given in the Spring, the difficulty of estimating the random error or random variation in the differences between tests is compounded. The error of measurement for a single test might actually operate in the case of one individual to lower his score in the Fall whereas in the Spring it might enlarge his score, thus making it seem as if he had made an enormous gain over the period of time in question. The opposite might be true also so that an individual who, in the truest sense of the word, had made normal progress throughout the period between tests might show up on the paired test scores as not having accomplished much of anything. The correlation between forms administered is also a factor here. If the same form is given over again the effects of practice compound the problem. We could elaborate in further detail but this is better done in a different context.

From the above it might be concluded that it is useless to test. Nothing could be farther from the truth. It is certainly very dangerous to use single paired comparisons (from Fall to Spring) for individuals as being fully true and dependable measures of what has happened to a child over a period of time between tests and within any one subject matter area. Only cumulative testing can do this.

Wherein Does the Error of Measurement Reside?

In this particular paragraph it is essential that we phrase our discussion in the form of a question to which no answer can ever be given in a completely definitive way. It is obvious from the discussion that has preceded this paragraph that the error of measurement may be in part due

to the inscrument itself. This kind of "error" is often estimated by computing a reliability coefficient for the instrument by correlating alternate halves of the test so that one collection or sequence of items constituting one-half the test is balanced by a second sequence of items as nearly as possible measuring the same thing. This is the so-called corrected split-half technique and, when the obtained correlations are corrected to allow for the full length of test, (Spearman-Brown Prophecy Formula) it does indeed give at least a rough estimate of the amount of stability characteristic of the test itself since the effect of the performance of any child is substantially ruled out by the fact that alternative items usually are taken within seconds of each other.

Other methods of computing the reliability of the test, however, call for the administration of the same form after an interval of time or the administration of two equivalent forms sequentially. In either case, there are additional difficulties involved.

In both cases the general perception of the test by the person taking it now becomes a factor. On one day an individual may be feeling fine and at his peak level of performance while on another day quite the opposite may be true. Failure to understand the directions, illness, fatigue, distraction, emotional upset, poor test administration, fear,--all of these factors may enter in to cause one test result to be different from another. When one studies test results for a group of students tested at an interval of seven months, as in this report, the personal factors relating to the pupils tested as contrasted to the factors embodied in the instrument itself may be overwhelmingly important especially if the pupil has had "a bad year". Indeed, there is a strong possibility that widely varying test scores for an individual especially over a short time span may be an indication of the emotional instability of that individual, especially when one test follows closely on the heels of the other.

The correlations between two tests administered seven months apart will always be lower than the correlation coefficient between paired scores on the test taken only days apart. One of several very important factors involved here is the amount of learning and forgetting that has taken place between the two test administrations, which alone would account for much of the difference in the results obtained from one time to the other e.g. fall to spring. This is not chance error subject to estimation by any formula. It is better known as "bias", i.e., the known and appraisable effect of what would generally be described as the experimental factors plus "static".

Perhaps this can best be seen if one considers a simple task, such as answering a single item, where the unreliability of the test question may be considered to be effectively minimized by the very nature of the item.

Consider the addition of three specified two-place numbers as the item in question using a free response mode of response. The answer will be right if the individual knows (1) the procedure for adding three two-place numbers, (2) knows his 100 addition facts, and (3) has the ability to retain in mind the partial sum of the first two numbers in the first column at the right while he adds to this partial sum the third number in the column, and, (4) if the sum of the first column is more than 9, ability to carry the remainder to the second column is also involved.

In other words, even in this apparently simple task of adding three two-place numbers there is a level of complexity that is not at all obvious on the surface. Any failure to remember a combination of two numbers, or any forgetting of the partial sum in the process of addition, etc., etc., results in an error that makes the final answer wrong. There are no partial credits. The second time the test is taken, especially if there has been drill and supplementary instruction in the task involved, the individual may have a better chance of answering the item correctly. On the other hand, if the item had been thought to have been effectively taught at the time it was first tested (according to the teacher's judgment) but was not touched on in the interim period, i.e., there was no maintenance of skills, forgetting would be a significant factor causing some individuals to have a higher potency for error the second time than the first.

It should be obvious from this illustration that it is not the test item that is at fault but it is something that actually happens to the individual child. This is bias,--not random error even though it does result in a lower correlation coefficient between Fall and Spring tests than between tests readministered within days.

Any test is, perforce, made up of a fairly large number of items and the test as a whole may or may not be homogeneous in the statistical sense of the word. Even an arithmetic computation test cannot truly be considered to be homogeneous because of the multiplicity of learnings involved. Even a test limited to addition alone cannot necessarily be considered to be entirely homogeneous. The one really homogeneous test in arithmetic (and this even might be challenged) would be a test involving knowledge of the hundred addition facts or the hundred subtraction facts.

When one moves into the field of reading it is obvious that the possibilities for errors increase since the content of the reading tests themselves is not material taken literally out of the body of instructional material, but represents a novel body of content on which the individual exercises his developed skill in reading materials suitable for his level of development whatever they may be. Even repeating the same paragraphs after a period of seven months, as was done in the New Hampshire situation, is no guarantee that the result on the second test is effectively reflecting

the outcome of instruction that has taken place between the first test and the second test, especially if the test material is (or was) not particularly well suited to the needs of the child on one or the other administration. Lack of local validity in the choice of paragraphs could seriously affect the score distribution. This cannot be charged off as error of measurement. Someone goofed! In a really good corrective program the correlation between first and second test may be lowered by the very effectiveness of the corrective instruction. A child having no good method of word attack on novel words initially but who improves greatly in this skill under instruction may make enormous gains on retest over a substantial period of time.

All of the above discussion boils down to one simple fact. When tests, made by fallible human beings, are administered by fallible human beings to fallible human beings chance variation and bias must be expected. All tests are short; the tests are imperfect as indicated by their reliability coefficients and by their correlations with valid criterion measures; but most important, the individuals taking the tests are variable in their performance from day to day to say nothing of their performance over a period of seven months.

Under these circumstances it only makes sense (1) to restrict one's broad interpretation of such data to general trends and (2) to identify the individuals who show extremely atypical performance over a period of time as the ones most likely to need special attention. In identifying these extreme cases, it does indeed make sense to keep in mind the standard error of the difference between scores since it then gives one confidence that an observed difference for an individual that is several times the standard error of the difference is most probably the one where some extraneous influences can be detected and analyzed.

SECTION VII
Part B
Comparisons in Raw Scores and Corresponding Grade Equivalents for the Random Sample and Title I Cases

In the previous pages we have discussed at length the hazards involved in making comparisons over a short period of time and the contribution of error measurement and bias in explaining differences that do occur.

We are now ready to look at the actual data for fall testing (October 1969) and spring testing (May 1970) for the random (representative) sample and for Title I cases.

In Section V we have already discussed the need for a random sample that would be representative of the state as a whole and we have documented the fact that this sample, selected carefully by sound statistical methods, was not in fact a random sample when it was actually used because of the failure of communities to test in accordance with the specifications. We have also established the fact that the tested random sample turned out to be quite representative of the statewide tested population. Therefore we feel that we are now ready to make actual score comparisons between fall and spring which will be valid.

In the first two tables presenting the data for fall and spring testing we have recorded the percentiles corresponding to the selected percentile ranks 75-50-25. We are not going to approach the interpretation of these tables so much from a statistical point of view as from a common sense point of view.

It must be remembered in this context that this may be the first time that this test-retest procedure involving a control sample has been employed in an operating situation; i.e., in a situation which was not part of an experimental research program. 1/ The intent was to find and test a sample which would be representative of our state and would therefore reveal what the goal should be for typical New Hampshire children in terms of gains on selected subtests of the Stanford Achievement Test. With this information available, obviously sounder judgments could be made about the performance of our Title I children.

This report has dealt extensively with the disadvantages and the inadequacies of grade equivalents as they are presently obtained and interpreted for measuring gain. It is for this reason that the gains have been listed first in raw score form, making possible a variety of interpretative procedures. The scores also have been interpreted in terms of the grade equivalents in order to make the pattern of interpretation fall into something comparable to the expected or traditional analysis as specified by the U.S. Office of Education Title I staff.

The median raw score gains have been circled in Tables VII-B-1 and VII-B-2 in order to make them stand apart from the other percentiles and to simplify the interpretation of these tables. As we look at the median of each test for the random samples in Grades 4 and 6, the first reaction is one of some consternation that the gains are as small as they are.

In these tables, as in some other tables in this report, the number of items in each test is given in parentheses in the margin alongside the test names. In making this evaluation, it is first necessary to make a value judgment or an assumption concerning the suitability of the test for the local curriculum. Keep in mind that the test items also have been arranged generally in order of in-

1/ Much credit must go to Mr. Richard Hodges, now State Director for Title I, for suggesting the procedure at the staff evaluation session following the 1968-69 testing program and implementing the procedure for the 1969-70 program.

Table VII-B-1

NEW HAMPSHIRE STATEWIDE TESTING PROGRAM 1969-70
PERCENTILES CORRESPONDING TO SELECTED PERCENTILE RANKS
WITH CORRESPONDING GRADE EQUIVALENTS FROM STANFORD GRADE NORMS
AND FROM TABLE OF EQUIVALENT METROPOLITAN NORMS

Random Sample

Grade 4

	Test	%ile Rank	Raw Score Fall	Raw Score Spring	Raw Score Gain	Stanford Intermediate I Grade Equiv. Fall	Stanford Intermediate I Grade Equiv. Spring	Stanford Intermediate I Grade Equiv. Gain	Stanford Intermediate I Grade Equiv. Dev.*	Comparable Metropolitan Grade Equiv. Fall	Comparable Metropolitan Grade Equiv. Spring	Comparable Metropolitan Grade Equiv. Gain	Comparable Metropolitan Grade Equiv. Dev.*
(38)	Word	75	21	27	6	4.9	5.9	1.0	+.3	5.2	6.4	1.2	+.5
	Meaning	50	15	22	(7)	3.9	5.1	1.2	+.5	4.2	5.4	1.2	+.5
		25	10	16	6	3.3	4.1	.8	+.1	3.6	4.4	.8	+.1
(60)	Paragraph	75	30	40	10	4.6	5.9	1.3	+.6	5.3	6.7	1.4	+.7
	Meaning	50	23	31	(8)	3.8	4.7	.9	+.2	4.4	5.4	1.0	+.3
		25	17	24	7	3.0	3.9	.9	+.2	3.3	4.5	1.2	+.5
(39)	Arithmetic	75	14	23	9	4.0	5.2	1.2	+.5	4.6	6.0	1.4	+.7
	Comp.	50	11	18	(7)	3.6	4.5	.9	+.2	4.1	5.2	1.1	+.4
		25	8	13	5	3.1	3.8	.7	0	3.5	4.3	.3	+.1
(32)	Arithmetic	75	16	20	4	4.8	5.5	.7	0	5.2	5.9	.7	0
	Concepts	50	12	16	(4)	4.1	4.8	.7	0	4.4	5.2	.8	+.1
		25	9	11	2	3.3	3.9	.6	-.1	3.5	4.2	.7	0
(32)	Arithmetic	75	16	21	5	4.6	5.5	.9	+.2	5.0	6.1	.5	-.2
	Appl.	50	12	16	(4)	4.0	4.6	.6	-.1	4.2	5.0	.8	+.1
		25	9	11	2	3.6	3.9	.3	-.4	3.8	4.1	.3	-.4

Grade 6

						Stanford Intermediate II				Comparable Metropolitan			
(48)	Word	75	32	36	4	7.3	8.0	.7	0	8.1	9.1	1.0	+.3
	Meaning	50	26	31	(5)	6.2	7.1	.9	+.2	6.8	7.8	1.0	+.3
		25	19	25	6	5.1	6.0	.9	+.2	5.5	6.5	1.0	+.3
(64)	Paragraph	75	40	47	7	6.7	7.8	1.1	+.4	7.6	9.2	1.6	+.9
	Meaning	50	33	39	(6)	5.9	6.6	.7	0	6.7	7.5	.8	+.1
		25	25	29	4	4.8	5.3	.5	-.2	5.5	6.0	.5	-.2
(39)	Arithmetic	75	17	23	6	5.9	6.8	.9	+.2	7.2	8.2	1.0	+.3
	Comp.	50	13	17	(4)	5.2	5.9	.7	0	6.0	7.2	1.2	+.5
		25	10	12	2	4.6	5.0	.4	-.3	5.4	5.8	.4	-.3
(32)	Arithmetic	75	16	21	5	6.5	7.6	.9	+.2	7.1	8.5	1.4	+.7
	Concepts	50	13	16	(3)	5.9	6.5	.6	-.1	6.4	7.1	.7	0
		25	9	11	2	4.9	5.4	.5	-.2	5.3	5.8	.5	-.2
(39)	Arithmetic	75	22	25	3	6.6	7.4	.8	+.1	7.4	8.2	.8	+.1
	Appl.	50	17	19	(2)	5.7	6.1	.4	-.3	6.4	6.9	.5	-.2
		25	12	13	1	4.6	4.9	.3	-.4	5.0	5.4	.4	-.3

* Represents the deviation from the expected gain of .7 of a calendar year, often inaccurately designated 7 months of a school year.

Table VII-B-2

NEW HAMPSHIRE STATEWIDE TESTING PROGRAM 1969-70
PERCENTILES CORRESPONDING TO SELECTED PERCENTILE RANKS
WITH CORRESPONDING GRADE EQUIVALENTS FROM STANFORD GRADE NORMS
AND FROM TABLE OF EQUIVALENT METROPOLITAN NORMS

Title I

Grade 4

	Test	%ile Rank	Stanford Intermediate I Raw Score Fall	Spring	Gain	Grade Equiv. Fall	Spring	Gain	Dev.*	Comparable Metropolitan Grade Equiv. Fall	Spring	Gain	Dev.*
(38)	Word Meaning	75	11	16	5	3.5	4.1	.6	-.1	3.8	4.4	.6	-.1
		50	8	12	(4)	3.1	3.6	.5	-.2	3.4	3.9	.5	-.2
		25	5	8	3	2.7	3.1	.3	-.4	3.0	3.4	.4	-.3
(60)	Paragraph Meaning	75	19	25	6	3.2	4.0	.8	+.1	3.6	4.6	1.0	+.3
		50	15	19	(4)	2.8	3.2	.4	-.3	3.1	3.6	.5	-.2
		25	12	15	3	2.5	2.8	.3	-.4	2.7	3.1	.4	-.3
(39)	Arithmetic Comp.	75	12	18	6	3.7	4.5	.8	+.1	4.2	5.2	1.0	+.3
		50	9	13	(4)	3.3	3.8	.5	-.2	3.7	4.3	.6	-.1
		25	6	10	4	2.7	3.5	.8	+.1	3.0	3.9	.9	+.2
(32)	Arithmetic Conc.	75	11	14	3	3.9	4.5	.6	-.1	4.2	4.8	.6	-.1
		50	8	10	(3)	3.0	3.6	.6	-.1	3.2	3.9	.7	.0
		25	6	8	2	2.5	3.0	.5	-.2	2.7	3.2	.5	-.2
(33)	Arithmetic Appl.	75	11	14	3	3.9	4.2	.2	-.5	4.1	4.5	.4	-.3
		50	8	10	(2)	3.4	3.8	.4	-.3	3.6	4.0	.4	-.3
		25	5	6	1	2.9	3.0	.1	-.6	3.1	3.2	.1	-.6

Grade 6

	Test	%ile Rank	Stanford Intermediate II Raw Score Fall	Spring	Gain	Grade Equiv. Fall	Spring	Gain	Dev.*	Comparable Metropolitan Grade Equiv. Fall	Spring	Gain	Dev.*
(48)	Word Meaning	75	21	27	6	5.4	6.4	1.0	+.3	5.8	7.0	1.2	+.5
		50	16	21	(5)	4.6	5.4	.8	+.1	4.9	5.8	.9	+.2
		25	12	16	4	3.9	4.6	.7	.0	4.2	4.9	.7	.0
(64)	Paragraph Meaning	75	27	34	7	5.0	6.0	1.0	+.3	5.7	6.8	1.1	+.4
		50	20	26	(6)	4.2	4.9	.7	.0	4.8	5.6	.8	+.1
		25	17	20	3	3.8	4.2	.4	-.3	4.4	4.8	.4	-.3
(39)	Arithmetic Comp.	75	14	17	3	5.4	5.9	.5	-.2	6.3	7.2	.9	+.2
		50	10	12	(2)	4.6	5.0	.4	-.3	5.4	5.8	.4	-.3
		25	7	9	2	3.8	4.4	.6	-.1	4.3	5.1	.8	+.1
(32)	Arithmetic Concepts	75	13	16	3	5.9	6.5	.6	-.1	6.4	7.1	.7	.0
		50	9	11	(2)	4.9	5.4	.5	-.2	5.3	5.8	.5	-.2
		25	7	8	1	4.3	4.6	.3	-.4	4.6	4.9	.3	-.4
(39)	Arithmetic Appl.	75	16	18	2	5.6	5.9	.3	-.4	6.3	6.7	.4	-.3
		50	12	13	(1)	4.6	4.9	.3	-.4	5.0	5.4	.4	-.3
		25	9	10	1	4.0	4.2	.2	-.5	4.2	4.5	.3	-.4

* Represents the Deviation from the Expected Gain of .7 of a calendar year, often inaccurately designated 7 months of a school year.

creasing difficulty, or in cycles of increasing difficulty within the subdivisions of the subtest content. Arranging the items in order of difficulty effectively counteracts any claim that the tests may have been too highly speeded. A less able child will do all he can do in the time allowed because the probability is great that the items beyond the point where he stops, assuming of course that he understands the directions, etc., will be too hard for him to answer correctly.

The evidence is clear from the score distributions that surprisingly few children guess wildly. For example, many times the median score of a distribution is below the chance level. If we are satisfied on these points, it leaves us with the necessity of asking if the number of points of gain shown in the first section of the tables is reasonable in terms of the number of items in each test.

We must assume that the Stanford Intermediate I test contained material substantially appropriate for us at the beginning of Grade 4. In Grade 4, the average scores earned in the arithmetic tests are on the low side in comparison with the number of items, but so is the performance of New Hampshire children according to Stanford norms.

In Grade 6, Intermediate II Battery, the New Hampshire median scores also tend to fall rather substantially below one-half the number of items in each of the three arithmetic tests. Before making any critical judgment at this point, it must be remembered that these batteries are intended to be suitable for two grades; namely, 4 and 5 for the Intermediate I, and 6 and 7 for the Intermediate II. Therefore, it is only right and proper that the number of items answered correctly should be somewhat less than half of the total number of items in the test in the lower of the two grades at each level.

More important than the median score at the beginning of the year is the amount of gain over the seven months between tests. Spring medians do go up appreciably, but do they go up enough? The amount of gain is more or less dependent upon the extent to which the content of the test is very specific to the instruction taking place during the period of time between first and second testing. Only an item-by-item subjective analysis of the test content by competent curriculum specialists will reveal to what extent the test items do measure the content of instruction.

In Table VII-B-3, the Stanford raw scores corresponding to a grade equivalent of 4.1 (October 15) and 4.8 (May 15) are tabled as nearly as these can be determined from the published norm tables for translating raw scores to grade equivalents.1/ Some fractional values have been given in this table because there were no precise corresponding scores given for 4.1 or 4.8 in the tables.

An examination of this table is very enlightening. Four plus points of gain are expected, according to the norms, in Word Meaning and about six in Paragraph Meaning at the Intermediate I level; also a five point score gain in Arithmetic Computation is stipulated. However, the expected gain for Concepts and Applications drops to four plus points. Note that these are the gains expected for the stipulated seven months, which is really .7 of a calendar year. 2/

In the Intermediate II Battery, very nearly comparable values are expected to result from seven months of in-school instruction between October and May.

These are not large gains. One would be much happier to have them at least half again as large. However, gain in score is not solely within the control of the test maker or publisher. Decreased emphasis on "book learning" with increased competition from other organized activities in school may be partly causative.

With these data in mind, let us go back to Table VII-B-1 and look to see what the students in New Hampshire did over the same length of time. In the seven months from the middle of October to the middle of May, the median for New Hampshire children in the random sample for Grade 4 reached or exceeded the amount of raw score and grade equivalent gains expected according to Stanford norms in all tests except Arithmetic Applications, where there was a .1 year deficit which, in part, is a smoothing effect.

In Grade 6 on the Intermediate II Battery, New Hampshire children in the random sample gained .2 year more than expected in Word Meaning, made the expected gain in Paragraph Meaning, and in Arithmetic Computation, had a minus .1 year deviation from the norm in Concepts, and were .3 year behind the expected gain in Arithmetic Applications.

Comparing Title I Gains with Expected Gains

It is always a problem to know what to expect of a group demonstrated ahead of time to be a less able group in terms of mental ability and known to come from the disadvantaged strata within the state. Although the generalization is somewhat dangerous, an examination of Table VII-B-2 compared to Table VII-B-1 suggests that the 75th percentile of the Title I children is not too far from the median value for the state as a whole.

1/ There is a question as to the appropriate norm to use because the tests were not all administered within the specified time limits in either fall or spring. 4.1 (or 6.1) versus 4.8 (or 6.8) seems suitable for comparison purposes.

2/ Differences between medians of successive grades are taken to represent the gain expected in a school year but are actually representative of a calendar year.

Table VII-B-3

Expected Change in Selected Stanford Subtest Scores
Over Seven Months of In-School Instruction

INTERMEDIATE I: GRADE 4

$SE_{Meas.}$ as Reported	Test	Raw Score Norm for 1/ October 15	May 15	Expected Gain
2.38	Word Meaning	16	20½	4½
3.10	Paragraph Meaning	26	32	6
2.36	Arithmetic Computation	15	20	5
2.40	Arithmetic Concepts	12	16	4
2.32	Arithmetic Applications	13	17½	4½

INTERMEDIATE II: GRADE 6

$SE_{Meas.}$ as Reported	Test	Raw Score Norm for 1/ October 15	May 15	Expected Gain
2.73	Word Meaning	25½	29½	4
3.22	Paragraph Meaning	35	40½	5½
2.41	Arithmetic Computation	18½	23	4½
2.51	Arithmetic Concepts	14	18	4
2.50	Arithmetic Applications	19	23	4

1/ Values read and interpolated from raw score-grade score tables given in accessory materials and handscoring booklets.

Table VII-B-4
New Hampshire Statewide Testing Program 1969-70
Fall and Spring Raw Score Means, Standard Deviations and Gains
Random Sample and Title I Cases

Grade 4

Part A - Random Sample

Test	No.of Items	Raw Score Means			Raw Score Standard Dev.	
		Fall	Spring	Gain	Fall	Spring
Word Mean.	38	15.92	21.87	5.95	7.10	7.26
Para. Mean.	60	24.44	31.97	7.53	9.43	10.50
Arith. Comp.	39	11.46	18.34	6.88	4.47	6.97
Arith. Conc.	32	12.88	16.37	3.49	5.20	6.06
Arith. Appl.	33	12.93	16.21	3.28	5.07	6.21
N = 585						

Part B - All Title I Cases

Word Mean.	38	9.13	13.21	4.08	4.98	6.11
Para. Mean.	60	16.35	20.85	4.50	5.97	7.93
Arith. Comp.	39	9.94	14.46	4.52	4.03	6.28
Arith. Conc.	32	9.37	11.71	2.34	4.41	5.17
Arith. Appl.	33	8.89	10.88	1.99	4.25	5.53
N = 434						

Grade 6

Part A - Random Sample

Word Mean.	48	25.67	30.07	4.40	8.49	8.22
Para. Mean.	64	32.91	38.31	5.40	11.23	12.21
Arith. Comp.	39	13.91	18.48	4.57	5.45	7.37
Arith. Conc.	32	13.52	16.53	3.01	5.10	6.49
Arith. Appl.	39	17.57	19.59	2.02	6.77	7.75
N = 645						

Part B - All Title I Cases

Word Mean.	48	17.45	22.23	4.78	7.32	8.22
Para. Mean.	64	22.56	28.16	5.60	9.17	10.60
Arith. Comp.	39	11.67	14.18	2.51	5.30	6.49
Arith. Conc.	32	10.36	12.70	2.34	4.50	5.61
Arith. Appl.	39	13.30	14.61	1.31	5.86	6.00
N = 235						

Table VII-B-5

A Comparison of Fall and Spring Gains
Involving the Median for Title I versus the
25th Percentile for the Random Sample

Grade 4
SAT: Intermediate I

	RANDOM SAMPLE RAW SCORE 25th PERCENTILE			TITLE I RAW SCORE MEDIAN		
	Fall	Spring	Gain	Fall	Spring	Gain
Word Meaning	10	16	6	8	12	4
Para. Meaning	17	24	7	15	19	4
Arith. Comp.	8	13	5	9	13	4
Arith. Conc.	9	11	2	8	10	2
Arith. Appl.	9	11	2	8	10	2

Grade 6
SAT: Intermediate II

	RANDOM SAMPLE RAW SCORE 25th PERCENTILE			TITLE I RAW SCORE MEDIAN		
	Fall	Spring	Gain	Fall	Spring	Gain
Word Meaning	19	25	6	16	21	5
Para. Meaning	25	29	4	20	26	6
Arith. Comp.	10	12	2	10	12	2
Arith. Conc.	9	11	2	9	11	2
Arith. Appl.	12	13	1	12	13	1

(In order to highlight this comparison, the pertinent information has been extracted from the above tables and is reproduced separately in Table VII-B-5.) Perhaps it is not unreasonable to say that the 25th percentile for the random sample constitutes a better goal for children in Title I than the state median does. For example, in Grade 4 of the Random Sample the Fall 25th percentile rank for Word Meaning is 10 points while the Fall median for Title I is 8 points. In Paragraph Meaning the Fall 25th percentile rank for the Random Sample (Grade 4) is 17 compared to the Title I median of 15 points. In Arithmetic Computation the Fall median for Title I is 9 and the 25th percentile rank for the Random Sample is 8. This comparison can be carried out too rigorously but it merely suggests a line of inquiry to the reader who examines and analyzes these data for himself.

In Table VII-B-1, comparable Metropolitan grade equivalents also are given. These were obtained from a table of equated grade equivalents for Metropolitan and Stanford provided by the publisher. By using the Stanford grade equivalent as the entry figure, it is possible to see what this grade equivalent would be in terms of Metropolitan '70 norms.

According to these new Metropolitan norms, New Hampshire children are doing definitely better than the national group was doing at the time Metropolitan was standardized in all subjects at the Grade 4 level and in Grade 6 in all but Concepts (normal gain) and Applications (deficit of .2 year).

Metropolitan norms have the advantage of being much more up-to-date than are Stanford norms but the interesting fact suggested by this analysis is that the net differences from fall to spring do not seem to be very different with one or two exceptions. Thus, it was harder to "make the grade" with Stanford norms but the net gains in grade equivalents from fall to spring aren't very different except in Computation in Grade 6 where Metropolitan norms seem to reflect a national downward trend. New Hampshire children make a net gain of 5 tenths according to Metropolitan norms while being just at grade on Stanford norms.

Summary

The picture arising out of the use of Metropolitan values stated to be equivalent to earned Stanford values is far more favorable to the State and in general must be said to be far more closely in line with what would be expected in terms of the mental ability of the children tested. Incidentally, it coincides far more closely, too, with previous results on former annual statewide testing programs at the 8th grade level where New Hampshire consistently has been at or near the national norm.

SECTION VIII
Bivariate Comparison of Fall-Spring Performance for the Random Sample and for Title I Cases

Part A
Bivariate Distributions as a Means of Comparing Fall and Spring Test Results

Whenever a test or series of tests is given at one period of time and repeated at a subsequent period, it is possible, of course, to study the stability of performance of the group as a whole in terms of the correlation between the variables, and at the same time, to study the extent to which an individual differs in his performance from the first period to the last by locating that individual on the bivariate distribution surface for the two variables. To put this in simpler language, it is possible to make a plot with the first test on one axis and the second test on the other axis, and from this plot work out the correlation coefficient giving the relationship between the two measures. The bivariate plot is not a necessary step to the computation of the correlation; it is more like a light to guide the aware person from accepting a statistically foolish result too often occurring due to hidden computational errors and to identify clusters of scores identifying pupils needing further checking.

There are certain conditions underlying the computation of Pearson product moment correlations relating to similarity of the shape of distribution on the two variables, etc., that are highly technical and need not be considered here. It suffices to say "for the record" that the relationship must be linear.

In this study the bivariate plots were made after the scores had been reduced to stanine form. It then was possible to see clearly the general level of relationship between the first test and the second. In the context of this study, this relationship is between the Fall test and the Spring test results separately by subtest. Since stanines are normalized standard scores, always having a mean of 5 and a standard deviation of 2 for scales based on the same population, such bivariates are especially easy to study.1/

The writer has determined empirically that the percent of cases falling within the mid-stanine range (a band three stanines wide running from lower left to upper right) will correspond almost exactly to the correlation coefficient if the stanines for the two variables are computed on the same group. The exception noted above makes this only approximately true for these distributions. The further significance of this finding about the relationship of the correlation coefficient and percent in the mid-stanine band will be discussed at a later point.

1/ There is a slight exception to the qualification "same population" in this study. Fall stanines are based on the total state group tested at each grade level; spring stanines are necessarily based on the performance of the random sample.

Statistical Correlation as a Process

When one studies the relationship between any two sets of data, using the Pearson product moment correlation procedure, the technique itself transforms the scores into standard scores with a mean of 0 and a standard deviation of 1. If raw scores are first transformed to standard scores such as stanines with the same mean and standard deviation, the correlation plot will be generally symmetrical and a line drawn on a bivariate chart from lower left to upper right will bisect the correlation surface. Normality of the distribution is not involved; symmetry is. The plotting of regression lines, i.e., lines drawn through the means of the arrays, will reflect the magnitude of the relationship existing between the two tests. If the correlation is plotted in terms of raw scores on the same test given twice, the difference between the means is roughly a measure of the average gain in raw scores that has been made by the group from the first testing period to the second. This does not mean that all individuals should have or even could show an equal gain.

No measure of gain is obvious when transformed scores, such as stanines, are used if these transformations are computed on the basis of scaling done independently for the two test administrations on the same group.

A child earning the same stanine both times has progressed as expected. An upward shift on the second test means an acceleration in his relative position in the group; a downward shift means less progress than would be true if he moved ahead at his expected rate.

When the tests being compared or correlated have been administered seven months apart, one must seek strenuously to find logical and persuasive reasons why some individuals perform poorly in one test and well in the other regardless of the order in which this difference occurs. Some part of the difference of course will be random error but not all and not more than would be true if the tests were administered within a short time span. Some part will be bias, i.e., changes resulting from identifiable causes. All identifiable factors influencing the performance of individuals must be diligently sought. Such differences as can be attributed to known influences are not assignable to the error of measurement!

For this situation Stsnford subtest scores for fall versus spring constitute the paired scores. Stanines were independently derived for fall and spring administrations. The position of any individual on any chart will reflect what has happened to that individual during the interim period but only in the sense that any change in his stanine means an upward or downward shift in his relative position in the group. Maintenance of his original status simply means that he has learned at the same rate as others like himself in ability.

It would be possible to study absolute gains in terms of standard scores only if the second test score is interpreted in terms of the standard scores assigned on the basis of the standard score transformation obtained from the distribution of the scores on the first test. This might have been done in the case of the New Hampshire data and, in some ways, it might have been more instructive than the procedure that was followed. 1/ Instead, as stated earlier, stanine comparisons are made in terms of fall stanines for the total state group and spring stanines for the random sample. The assignment of spring stanines on the basis of the random (representative) sample scores was a necessary condition; it was basic to the whole idea of testing the random (representative) sample in the spring in order to provide some reasonable way of comparing fall-spring performance.

Knowledge of the existence and significance of the regression effect for all individuals in the bivariate distribution except those at the mean further helps one in his attempt to make sense in the interpretation of paired comparisons. "Regression" is the name for a phenomenon widely ignored or misunderstood, namely, the tendency for high first measures to be lower on the second measure on a comparable instrument and vice versa. Tall parents have tall children but not as tall as as they are and vice versa. This effect is always present. Low scoring individuals tend to improve on a second test just by chance; high scorers tend to fall back. Only when the shift is greater than can reasonably be accounted for by chance can one be sure the shift is due to a systematic influence.

A little exercise of common sense after chasing down the protocols for deviant individuals so as to study the performance of these individuals from one time to another often can turn up the logical reasons why a particular performance was so atypical.

For example, perhaps on one test answer sheet the marks were not sufficiently heavy for the optical scanner to pick them up satisfactorily, while on another test the marks were quite readable. This would, of course, invalidate the two test comparisons for that individual.

Perhaps on one occasion the individual might have guessed substantially, marking every item on the test "with his eyes closed", so to speak, after he had done as much as he could do in terms of his own knowledge. This guessing factor on a test made up of four or five multiple choice questions would substantially raise his score and therefore his stanine placement. If, on the second test, he had sufficient self-confidence or he had been taught in the meantime that testing is supposed to be a true communication act calling for truthful responses of a non-chance nature and that he only does himself harm by guessing, his score the second time might sctually be lower than it was the first but would be more truly reflective of his status in the group.

With these thoughts in mind, it will be most helpful and provocative to study the following bivariate charts. (See Charts VIII-1 to VIII-10.)

1/ A separate study is under consideration.

Chart VIII-1

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate I Battery: Form X

Grade 4

RANDOM SAMPLE

Word Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9							9	6	6	21
8				2	2	7	11	12	10	44
7			1	3	12	13	31	6	8	74
6	1	1	3	19	23	21	22	4	1	95
5	2	1	8	28	17	22	12	3	1	124
4	3	4	25	32	25	10	3			102
3	8	12	24	15	5	1				65
2	10	4	16	7	2					39
1	4	5	7	1	1	1				19
Stanine	28	27	84	107	117	75	88	31	26	583

r = .75

Fall Mean = 5.05 Standard Deviation = 1.9

Spring Mean = 4.96 Standard Deviation = 2.0

% in Mid-Stanine Band = .74

TITLE I

Word Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9								2		2
8	1									1
7				1	1	2	4	1		9
6	2			7	10	1	2			22
5	2	2	9	10	6	2	1			32
4	20	11	42	17	10	2				102
3	46	18	31	18	5	1				119
2	31	12	18	7	1				1	70
1	35	9	22	4		1				71
Stanine	137	52	122	64	33	9	7	3	1	428

r = .52

Fall Mean = 3.17 Standard Deviation = 1.5

Spring Mean = 2.71 Standard Deviation = 1.6

Chart VIII-2

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate I Battery: Form X

Grade 4

RANDOM SAMPLE

Paragraph Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9					1	2	3	5	15	26
8		1			1	3	15	11	5	36
7				5	5	24	22	13	4	73
6	1	2	5	13	29	24	18	4	1	97
5	4	5	14	37	35	22	7	1	1	126
4	5	5	26	30	20	13	4	1		104
3	3	9	11	17	11	3	4	1		59
2	7	6	14	8	2	4	1			42
1	3	6	8		2		1	1		21
Stanine ↑	23	34	78	110	106	95	75	37	26	584

r = .67

Fall Mean = 5.03 Standard Deviation = 1.9
Spring Mean = 5.00 Standard Deviation = 2.0
% in Mid-Stanine Band = .73

TITLE I

Paragraph Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9										0
8	1								1	2
7			1		1	1	2	1		6
6	1			5	3	2	2			13
5	7	4	20	17	20	3	2			73
4	23	13	39	16	7	1	1			100
3	25	22	34	20	5		1			107
2	20	20	21	9	9	1				80
1	14	12	16	4	2	1				49
Stanine ↑	91	71	131	71	47	9	8	1	1	430

r = .37

Fall Mean = 3.32 Standard Deviation = 1.4
Spring Mean = 2.95 Standard Deviation = 1.5

Chart VIII-3

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate I Battery: Form X

Grade 4

RANDOM SAMPLE

Arithmetic Computation

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ⟵
9				1	2		8	4	10	25
8					5	7	8	5	6	31
7		1	4	8	11	24	24	10	3	85
6	2		1	23	14	15	12	4	1	72
5	3	7	24	33	36	35	16	4	4	162
4	7	2	9	15	10	8	6	1		58
3	7	6	13	22	9	12	4	1	1	75
2	6	9	9	11	9	6		1		51
1	7	2	8	6			1			24
Stanine ↑	32	27	68	119	96	107	79	30	25	583

r = .56

Fall Mean = 4.96 Standard Deviation = 2.0
Spring Mean = 4.98 Standard Deviation = 2.0
% in Mid-Stanine Band = .59

TITLE I

Arithmetic Computation

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ⟵
9					3	2	1			6
8			2	4		1	3	3	2	15
7		1	2	14	8	9	2	3	3	42
6	5	3	4	12	8	9	3			44
5	11	5	18	27	17	11	11	2		102
4	4	5	9	15	4	4	1			42
3	14	10	25	21	8	3	3			84
2	14	4	17	13	11	5	1	1		66
1	10	4	10	3			1			28
Stanine ↑	58	32	87	109	59	44	26	9	5	429

r = .43

Fall Mean = 4.24 Standard Deviation = 1.9
Spring Mean = 3.90 Standard Deviation = 1.9

Chart VIII-4

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate I Battery: Form X

Grade 4

RANDOM SAMPLE

Arithmetic Concepts

Spring

Fall	1	2	3	4	5	6	7	8	9	Stanine ←
9			1			1	2	3	14	21
8						5	14	13	6	38
7		1	1	1	10	25	15	9	4	66
6		1	7	11	16	33	19	7	2	96
5	3	3	13	21	25	37	4	1		107
4	7	9	22	32	39	14	2	4		129
3	6	11	20	10	9	11	1			68
2	3	4	7	7	4	2				27
1	3	6	12	5	3					29
Stanine ↑	22	35	83	87	106	128	57	37	26	581

r = .68

Fall Mean = 4.93 Standard Deviation = 1.9
Spring Mean = 5.01 Standard Deviation = 1.9
% in Mid-Stanine Band = .70

TITLE I

Arithmetic Concepts

Spring

Fall	1	2	3	4	5	6	7	8	9	Stanine ←
9						1	1		2	4
8	1							1		2
7		1	2		2	9	4	2		20
6	1	5	3	7	14	9	5	1		45
5	2	3	10	9	6	9	1	1		41
4	13	21	30	26	9	14	2		1	116
3	8	16	26	23	10	4				87
2	9	9	14	6	5	1				44
1	20	20	26	8	2					76
Stanine ↑	54	75	111	79	48	47	13	5	3	435

r = .53

Fall Mean = 3.57 Standard Deviation = 1.8
Spring Mean = 3.52 Standard Deviation = 1.7

Chart VIII-5

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate I Battery: Form X

Grade 4

RANDOM SAMPLE

Arithmetic Applications

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9					1	1	2	7	9	20
8		1		1	1	11	13	6	9	44
7	1	1	1		3	12	19	6	3	46
6	1	2	7	13	16	32	26	6	6	109
5	1	6	6	22	25	35	9	2	1	107
4	8	10	20	44	35	15	4	1		137
3	6	11	9	17	10	6	1	1		61
2	7	8	8	6	7	1	1			38
1	5	3	3	4	1	1				17
Stanine	29	42	54	107	99	114	77	29	28	579

r = .65

Fall Mean = 4.95 Standard Deviation = 1.9
Spring Mean = 5.01 Standard Deviation = 2.0
% in Mid-Stanine Band = .70

TITLE I

Arithmetic Applications

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9							3		1	4
8				2		1				3
7	1					5	1		1	8
6		4		6	5	8	7	2		32
5	6	7	8	13	13	8	4	1		60
4	16	15	15	21	20	4	2			99
3	21	21	17	22	6	2		1		90
2	23	19	13	16	2	2				75
1	23	14	6	14	1					58
Stanine	90	80	59	100	47	30	17	4	2	429

r = .53

Fall Mean = 3.45 Standard Deviation = 1.7
Spring Mean = 3.28 Standard Deviation = 1.8

Chart VIII-6

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate II Battery: Form X

Grade 6

RANDOM SAMPLE

Word Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9							7	4	16	27
8						5	17	9	9	40
7				2	5	22	41	9	4	83
6			1	10	35	26	25	3	1	101
5		2	12	38	52	30	8	1		143
4	1	5	21	34	19	14			1	95
3	4	16	24	26	6	5				81
2	7	18	14	5	1					45
1	11	8	5	1		1				26
Stanine →	23	49	77	116	118	103	98	26	31	641

r = .83

Fall Mean = 4.99 Standard Deviation = 2.0
Spring Mean = 4.98 Standard Deviation = 2.0
% in Mid-Stanine Band = .83

TITLE I

Word Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9							1	1	1	3
8							1		1	2
7						1	1	1		3
6				3		1	2		1	7
5			5	6	1	6	2			29
4		8	10	11	4	1	2			36
3	6	13	24	19	3	2				67
2	10	15	19	4	1					49
1	19	8	7			1				35
Stanine →	35	44	65	43	18	12	9	2	3	231

r = .74

Fall Mean = 3.15 Standard Deviation = 1.6
Spring Mean = 3.28 Standard Deviation = 1.7

Chart VIII-7

Stanine Bi .riate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate II Battery: Form X

Grade 6

RANDOM SAMPLE

Paragraph Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9							7	9	14	30
8					1	6	23	9	5	44
7	1		3	1	7	25	25	10	5	77
6				4	35	36	13	8	2	98
5	2	1	14	35	46	30	12	1	1	142
4	4	4	22	29	29	14	1			103
3	5	19	17	27	11	3				82
2	6	13	20	3						42
1	6	8	8	1	1					24
Stanine →	24	45	84	100	130	114	81	37	27	642

r = .79

Fall Mean = 5.02 Standard Deviation = 2.0

Spring Mean = 4.99 Standard Deviation = 1.9

% in Mid-Stanine Band = .80

TITLE I

Paragraph Meaning

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine ←
9							1		2	3
8							1			1
7			1		1	3	1	1		7
6					1	5		1		7
5		1	1	10	10	6				28
4		3	9	14	11	1	1			39
3	10	13	19	16	11	1				70
2	12	14	22	4	1					53
1	6	11	7	5						29
Stanine →	28	42	59	49	35	16	4	2	2	237

r = .71

Fall Mean = 3.23 Standard Deviation = 1.6

Spring Mean = 3.45 Standard Deviation = 1.6

Chart VIII-8

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate II Battery: Form X

Grade 6

RANDOM SAMPLE

Arithmetic Computation

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9						5	5	11	10	31
8		1		2	3	5	8	6	8	33
7			2	6	12	28	25	11	5	89
6	2	1	3	16	31	31	22	5	2	113
5	1	5	8	19	26	29	11	1		100
4	5	10	45	26	34	16	7	2		145
3	4	9	22	13	15	4	3			70
2	7	3	10	9	4	1	1	1		36
1	8	5	9	6		1				29
Stanine →	27	34	99	97	125	120	82	37	25	646

r = .67

Fall Mean = 5.00 Standard Deviation = 1.9
Spring Mean = 4.98 Standard Deviation = 1.9
% in Mid-Stanine Band = .68

TITLE I

Arithmetic Computation

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9						1	1		1	3
8				1		1	5	2		9
7			1	3	8	7	3	2	1	25
6		1	3	6	5	3	5			23
5	2	4	8	2	9	4				29
4	3	9	13	12	9	5	1			52
3	5	9	14	10	5		1			44
2	7	4	10	4	2	2				29
1	9	4	5	3						21
Stanine →	26	31	54	41	38	23	16	4	2	235

r = .64

Fall Mean = 4.15 Standard Deviation = 2.0
Spring Mean = 3.84 Standard Deviation = 1.8

Chart VIII-9

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate II Battery: Form X

Grade 6

RANDOM SAMPLE

Arithmetic Concepts

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9						1	7	7	11	26
8					4	5	13	11	8	41
7				3	15	25	17	7	7	74
6	2	3	3	13	32	44	22	3	4	126
5		5	14	28	34	31	14	1		127
4	4	10	11	23	27	12	2			89
3	8	15	20	30	12	5	1			91
2	8	14	8	11	2	1				44
1	4	7	7	8	1					27
Stanine	26	54	63	116	127	124	76	29	30	645

r = .73

Fall Mean = 4.98 Standard Deviation = 2.0

Spring Mean = 4.96 Standard Deviation = 1.9

% in Mid-Stanine Band = .71

TITLE I

Arithmetic Concepts

Fall \ Spring	1	2	3	4	5	6	7	8	9	Stanine
9				1		1		1	1	4
8							2		1	3
7					1	3	6			10
6		1		5	4	7	3			20
5	1	2	4	22	13	6	2			50
4	4	5	5	11	7	4				36
3	4	11	12	17	6	1				51
2	7	9	5	5	4	1				31
1	9	9	7	3	2	1				31
Stanine	25	37	33	64	37	24	13	1	2	236

r = .65

Fall Mean = 3.77 Standard Deviation = 1.9

Spring Mean = 3.81 Standard Deviation = 1.7

Chart VIII-10

Stanine Bivariate Charts Showing the Relationship between Fall and Spring Results for Selected Stanford Achievement Subtests Given in School Year 1969-70 Intermediate II Battery: Form X

Grade 6

RANDOM SAMPLE

Arithmetic Applications

Spring

Fall	1	2	3	4	5	6	7	8	9	Stanine ←
9					1		8	12	7	28
8						3	6	12	10	31
7				3	12	33	31	14	4	97
6	2	4	3	8	14	36	11	8		88
5	3	4	14	42	51	35	9	1		159
4	3	10	14	53	20	13	1			114
3	7	3	16	9	7	1				43
2	9	9	12	15	7	2				54
1	9	4	7	8						28
Stanine →	33	34	66	138	112	125	66	47	21	642

r = .74

Fall Mean = 5.02 Standard Deviation = 1.9

Spring Mean = 4.96 Standard Deviation = 1.9

% in Mid-Stanine Band = .74

TITLE I

Arithmetic Applications

Spring

Fall	1	2	3	4	5	6	7	8	9	Stanine ←
9							1	1		2
8							2			2
7				1	2	8	4	1		16
6	3	3	1	1	8	2				18
5	1	1	8	15	13	9				47
4	1	5	10	26	9	3				54
3	5	2	8	13	6	1				35
2	4	6	7	9	2					28
1	9	7	13	3						32
Stanine →	23	24	47	68	40	23	7	2	0	234

r = .61

Fall Mean = 3.83 Standard Deviation = 1.8

Spring Mean = 3.79 Standard Deviation = 1.6

SECTION VIII
Part B
Data Concerning the Measures of Relationship
Between Tests Administered in the Fall
And Repeated in the Spring

Immediately preceding this discussion the writer has made an attempt to deal with a few of the issues involved in the interpretation of correlation coefficients. Much more could be said but even these cautionary notes may be considered by some readers to be superfluous. Essentially, the task of making sense out of correlation coefficients calls for a level of statistical competence and sophistication that probably does not characterize more than a small fraction of the people in public education at both administrative and instructional levels.

Few people, for example, are aware of the fact that reliability coefficients can be unduly inflated by drawing a sample that is as heterogeneous as possible. As a matter of fact, the population samples used for determining reliability coefficients for the Stanford consisting of 1000-case random samples from the standardization group probably are about as variable as any group could be and remain within a grade. Sensible reliability coefficients are computed on well described community samples so that the values obtained will be descriptive of the local scene. Since we have no data on reliability for the communities within this State, we are including in Table VIII-B-1 the split-half reliability coefficients for Stanford subtests used in this study as reported in the Technical Manual. These values almost surely overestimate the tests' reliability in the context in which they are used, but lacking something better, they will have to serve the purpose.

In the next adjacent column, the correlations between fall and spring tests are reported for each of the five Stanford tests consistently studied in this report. The correlations indicate the relationship between the Stanford subtests given in October of '69 and repeated in May of '70. The data are given separately for Grade 4 and Grade 6 and for the random (representative) sample and Title I. Notice that the correlations are consistently lower in Grade 4 than in Grade 6. The writer knows of no systematic and underlying cause for the difference in the magnitude of these values.

Possibly the subtest scores for the Intermediate I Battery were less normally distributed than those for the Intermediate II Battery which was used at Grade 6. Possibly these tests were more relevant to the instruction in the 6th grade than at the 4th grade. Or perhaps the bias in the two grade populations is a sufficient explanation. This inconsistency is only an illustration of the fact that correlation coefficients are not self-interpreting statistics with a common meaning regardless of the situation within which they were obtained. All correlations for the random sample are lower by a substantial margin than the reported reliability coefficients. This is really not unexpected since the reported reliabilities are instrument reliabilities (pupil variation controlled) and contrasted to the test-retest comparisons.

When one moves to similar values for Title I, it is interesting to note that in every instance, with one possible exception in Grade 6, the correlations are clearly lower than for the random sample and lower in the spring than in the fall. Above all else, this reflects the fact that the Title I group is substantially less variable than the random sample. However, we would like to think that some part of the lowering of the correlations actually is due to something that happened to the children during the period from October to May. If pupils have validly diagnosed remedial defects and if they are provided with adequate instruction to counteract specifically the defined and described defects, the net result would be to lower the correlation between their first testing and their second due to the fact that the amount of gain or improvement under special instruction will vary widely among individuals, depending in large measure on the extent or magnitude of their difficulty in the first place, and the effectiveness of the special instruction. For example, children known to have a correctible reading deficiency based upon adequate diagnosis will improve greatly over a relatively short period of time, but certainly not in a manner consistent from individual to individual since this depends upon the nature of the defect in the first place and the adequacy of the instruction to correct it in the second place.

Tables VIII-B-2 and 3 reproduce eight correlation matrices showing the intercorrelations of the Stanford subtests separately for the random sample and for the Title I cases and separately for Fall and Spring for Grades 4 and 6. These eight matrices are interesting indeed to study but are only a basis for speculation without knowing a lot more about what took place between the Fall testing program and the Spring follow-up program. We must assume that the program of instruction for the children in the random sample was just about normal or typical of what goes on ordinarily. That being the case, it is rather clear that even here influences and factors are at work which tend to dilute the degree of agreement between a series of tests taken two at a time.

Perhaps this a good place to leave to those who wish to speculate as to the significance of these mathematical coefficients what their significance in truth may be and turn, for the benefit of those who are perhaps more visually minded and less statistically oriented, to a consideration of the small bivariate distributions set up in terms of stanines which make more evident the nature of the relationship between Fall and Spring results separately by tests and separately for the random sample compared to the Title I group.

The first series of bivariate charts relates to Grade 4 and each test's bivariate chart for the random sample is paired with the corresponding bi-

variate chart for that test for Title I children.

As one goes from chart to chart, it is evident that there are some maverick cases in almost every chart where it seems quite unlikely that the protocols, i.e., that these individual test results, were valid in both instances. For example, in the Paragraph Meaning bivariate chart for the random sample, it is evident that a child who earned a stanine of 8 in the Fall would be most unlikely to earn a valid stanine of only 2 when retested in the Spring. Every peculiar case of this kind, falling far out from the general cluster of scores should have been investigated case by case.

However the general practice is NOT EVEN TO HAVE THE ANSWER DOCUMENTS RETURNED. The least valuable part of the total information obtained by testing is given complete primacy. Having item analysis information is no help. Having a chart showing how every item was answered is better but none of these helps one learn why some pupils acted erratically and none permits the pupil to share adequately with the teacher his areas of strength and weakness.

These data are available for further study since the answer sheets or scoreable booklets were returned and have been stored in the hope that funds could be made available for making this kind of detailed inspection for the sake of what it might show up by way of insights into the dynamics of testing and thus improve future programs.

Generally, however, the frequencies are clustered, more or less symmetrically, centered around the mid-stanine range or band. It must be remembered that in these bivariates, Fall stanines were based upon the total sample of children tested throughout the state ranging from 11,700+ in Grade 6 to 12,000+ in Grade 4 because these stanines were already on the tape. Stanines based upon the representative sample might have been preferable but the others already were in the hands of the local schools. The Spring stanines, on the other hand, are based, of necessity, upon the performance of the tested random sample. These stanines were used, of course, to interpret the results for Title I cases also.

The correlation coefficients are given at the bottom of each of the bivariate charts and for the random sample only, the percent of children in the mid-stanine band is also reported. It will be observed that this percentage is very similar to the correlation coefficient and if both sets of data, Spring versus Fall, had been based upon the same stanine tranaformation, these percentages would be even closer. Similar percentages are not given for Title I because the stanines were not independently derived for that sample and the advantages of this comparison with r would have been lost.

Generally speaking, there is a curtailment in the distribution for Title I which shows up by a thinning of the scatterplot in the upper right-hand corner. This is most evident if one looks at the marginal figures and notes that the distributions are skewed in the sense that there are fewer cases in the upper ranges for Title I than for the random sample which are more or less symmetrical.

Since the stanines for Fall and the stanines for Spring were computed independently, one cannot observe growth directly by comparing the Fall and Spring performance. This point has been discussed earlier in the introductory portions of this section. It will be noted, however, that the random sample means closely approximate 5 and the random sample standard deviations closely approximate 2 in every bivariate for both grades. On the other hand, the means for the Title I sample tend to be substantially lower, especially in the Spring and while the standard deviations vary, they also tend to be somewhat smaller than those for the random sample. 1/

Considering just the Title I pupils, it is evident that the relationship, i.e., correlation, is far from 1.00 between the Fall and Spring data but it is also evident that it would be possible to identify children falling substantially outside the mid-stanine range who should have been investigated pupil by pupil if these data had been reported promptly enough to the schools. The ideal arrangement would have been to have the answer sheets for all children returned to each school and for someone to undertake the task at the local level of examining suspect answer sheets in terms of the Fall-Spring paired responses to see which responses failed to be consistent from one testing program to another. This would have been especially helpful in this instance since even the same FORM was used.

The bivariates deserving most serious study are those relating to Paragraph Meaning since this was the curriculum area where major emphasis was put in the Title I program. However, in doing so, please remember that these data describe all Title I children, not just those in reading programs. A separate analysis will be prepared as a supplement to this report at a later date analyzing the data in a somewhat similar fashion for those who were in remedial reading programs.

Bivariate charts of this sort take on their greatest importance as a basis for helping a teacher, administrator, or supervisor to identify individual cases and study them against the background of the performance of the group as a whole. An isolated case is hard to interpret; a case in a defined and charted distribution is more easily studied and understood. For this reason, it is also most helpful to have the data for the typical or random sample for comparison with the specifically designated Title I cases. It is also possible that even the sophisticated will have a renewed sense of what a correlation really means if those who bother to read the report take a good look at the charts.

1/ See Table VIII-B-4.

With the presentation of the data on these bivariate charts, the statistical portions of this report are completed. It just remains, therefore, to sum up and to provide the reader with the writer's own evaluation of the total program in terms of an overview of all of the data available. Obviously, this is a highly subjective process and disagreement as to the significance of these data can be expected. Every individual in a position of responsibility must perform this tedious task of studying the data for himself. It has been this writer's intention within the sadly lacking basic data to select and highlight those parts that to him seemed most significant.

Table VIII-B-1

Correlations* Between Selected Stanford Subtests Administered in the Fall and Repeated in the Spring In Comparison with Reported Reliability Coefficients

Grade 4 - 1969-70
SAT: Int. I: X

	r_{1I}**	Random Sample	Title I
Word Meaning	.90	.77	.56
Paragraph Meaning	.92	.69	.43
Arithmetic Computation	.89	.60	.46
Arithmetic Concepts	.86	.71	.57
Arithmetic Applications	.86	.68	.57

Grade 6 - 1969-70
SAT: Int. II: X

	r_{1I}**	Random Sample	Title I
Word Meaning	.90	.83	.75
Paragraph Meaning	.93	.82	.72
Arithmetic Computation	.89	.69	.70
Arithmetic Concepts	.85	.76	.69
Arithmetic Applications	.89	.79	.63

* Based on raw scores.
** Corrected split half reliability coefficient as reported by the publisher, based on random samples of 1,000 cases per grade from the standardized sample.

Table VIII-B-2

Intercorrelations* of Selected Stanford Subtests
for Random Sample and for Title I Separately Tested
Fall and Spring

Grade 4: FALL
SAT: Intermediate I

A
RANDOM SAMPLE

Test Name and Number		1	2	6	7	8
Word Mng.	1	1.00				
Para. Mng.	2	.72	1.00			
Arith. Comp.	6	.29	.38	1.00		
Arith. Conc.	7	.54	.57	.48	1.00	
Arith. Appl.	8	.52	.53	.47	.72	1.00

B
TITLE I

Test Name and Number		1	2	6	7	8
Word Mng.	1	1.00				
Para. Mng.	2	.56	1.00			
Arith. Comp.	6	.23	.33	1.00		
Arith. Conc.	7	.37	.48	.46	1.00	
Arith. Appl.	8	.39	.44	.35	.53	1.00

Grade 6: FALL
SAT: Intermediate II

C
RANDOM SAMPLE

Test Name and Number		1	2	5	6	7
Word Mng.	1	1.00				
Para. Mng.	2	.80	1.00			
Arith. Comp.	5	.47	.52	1.00		
Arith. Conc.	6	.62	.61	.59	1.00	
Arith. Appl.	7	.65	.66	.62	.76	1.00

D
TITLE I

Test Name and Number		1	2	5	6	7
Word Mng.	1	1.00				
Para. Mng.	2	.76	1.00			
Arith. Comp.	5	.44	.54	1.00		
Arith. Conc.	6	.59	.57	.57	1.00	
Arith. Appl.	7	.58	.63	.57	.68	1.00

*These correlations, which are based on raw scores, tend to run 2 to 3 points lower than the stanine correlations reported with the bivariates, due to coarseness of grouping in the case of stanines.

Table VIII-B-3

Intercorrelations of Selected Stanford Subtests
for Random Sample and for Title I Separately Tested
Fall and Spring

Grade 4: SPRING
SAT: Intermediate I

A
RANDOM SAMPLE

Test Name and Number		1	2	6	7	8
Word Mng.	1	1.00				
Para. Mng.	2	.77	1.00			
Arith. Comp.	6	.38	.47	1.00		
Arith. Conc.	7	.60	.62	.58	1.00	
Arith. Appl.	8	.58	.66	.53	.75	1.00

B
TITLE I

Test Name and Number		1	2	6	7	8
Word Mng.	1	1.00				
Para. Mng.	2	.61	1.00			
Arith. Comp.	6	.33	.42	1.00		
Arith. Conc.	7	.42	.53	.54	1.00	
Arith. Appl.	8	.46	.56	.56	.69	1.00

Grade 6: SPRING
SAT: Intermediate II

C
RANDOM SAMPLE

Test Name and Number		1	2	5	6	7
Word Mng.	1	1.00				
Para. Mng.	2	.80	1.00			
Arith. Comp.	5	.48	.61	1.00		
Arith. Conc.	6	.62	.70	.71	1.00	
Arith. Appl.	7	.64	.71	.67	.79	1.00

D
TITLE I

Test Name and Number		1	2	5	6	7
Word Mng.	1	1.00				
Para. Mng.	2	.73	1.00			
Arith. Comp.	5	.44	.52	1.00		
Arith. Conc.	6	.56	.59	.64	1.00	
Arith. Appl.	7	.61	.62	.62	.73	1.00

Table VIII-B-4
Means, Standard Deviations, and Correlation Coefficients
for the Random Sample and Title I Cases, Fall versus Spring

Grade 4

SAT Int. I Subtest	RANDOM SAMPLE Mean	Standard Deviation	r		TITLE I Mean	Standard Deviation	r
Word Meaning				Word Meaning			
Spring	4.96	2.0	.75	Spring	2.71	1.5	.52
Fall	5.05	1.9		Fall	3.17	1.6	
Difference	-.09		N = 583	Difference	-.46		N = 428
Para. Meaning				Para. Meaning			
Spring	5.00	2.0	.67	Spring	2.95	1.5	.37
Fall	5.03	1.9		Fall	3.32	1.4	
Difference	-.03		N = 584	Difference	-.37		N = 430
Arith. Comp.				Arith. Comp.			
Spring	4.98	2.0	.56	Spring	3.90	1.9	.43
Fall	4.96	2.0		Fall	4.24	1.9	
Difference	.02		N = 583	Difference	-.34		N = 429
Arith. Conc.				Arith. Conc.			
Spring	5.01	1.9	.68	Spring	3.52	1.7	.53
Fall	4.93	1.9		Fall	3.57	1.8	
Difference	.08		N = 581	Difference	-.05		N = 435
Arith. Appl.				Arith. Appl.			
Spring	5.01	2.0	.65	Spring	3.28	1.8	.53
Fall	4.95	1.9		Fall	3.45	1.7	
Difference	.06		N = 579	Difference	-.17		N = 429

Grade 6

SAT: Int. II							
Word Meaning				Word Meaning			
Spring	4.98	2.0	.83	Spring	3.28	1.7	.74
Fall	4.99	2.0		Fall	3.15	1.6	
Difference	-.01		N = 641	Difference	.13		N = 231
Para. Meaning				Para. Meaning			
Spring	4.99	1.9	.79	Spring	3.45	1.6	.71
Fall	5.02	2.0		Fall	3.23	1.6	
Difference	-.03		N = 642	Difference	.22		N = 237
Arith. Comp.				Arith. Comp.			
Spring	4.98	1.9	.67	Spring	3.84	1.8	.64
Fall	5.00	1.9		Fall	4.15	2.0	
Difference	-.02		N = 646	Difference	-.31		N = 235
Arith. Conc.				Arith. Conc.			
Spring	4.96	1.9	.73	Spring	3.81	1.7	.65
Fall	4.98	2.0		Fall	3.77	1.9	
Difference	-.02		N = 645	Difference	.04		N = 236
Arith. Appl.				Arith. Appl.			
Spring	4.96	1.9	.74	Spring	3.79	1.6	.61
Fall	5.02	1.9		Fall	3.83	1.8	
Difference	-.06		N = 642	Difference	-.04		N = 234

SECTION IX

A Personal Commentary

I feel it quite necessary at this point to evaluate this report and especially all that has led up to it in terms of what lessons it may have taught as well as what it "proves" about Title I programs in New Hampshire and, by implication, elsewhere.

1. The process has been too time-consuming by many months. This has resulted because of lack of coordination at all levels and between all agencies involved. To a very large extent this was inevitable at first as we groped our way toward a configuration that would answer our questions about the effectiveness of Title I programs and still deal realistically with the mensuration problems involved. Before-after testing with instruments built for a different purpose has a built-in "bomb" in the reality of errors of measurement enhanced when two fallible measures are compared over a seven-month time span.

There are no precedents to follow, and any shallow or superficial analysis using inappropriate sampling statistics will neither reveal the inherent dangers nor provide insightful suggestions for the future.

My own conclusion is that the variations in individual pupil performance noted on our tests are quite as much due to the pupil's built-in day-by-day variability plus the inefficiencies of our present educational process as they are to the instruments themselves.

I have tried to bring this out in several places in the text. Better instruments are needed, to be sure, but no instrumentation no matter how good, will nullify variability built into the situation, not the tests.

2. New Hampshire is a remarkably typical state as determined by national norms. The use of the MAT equivalence tables to re-interpret Stanford data reinforces the Otis-Lennon data in establishing this conclusion which I have repeatedly observed over the last twenty years.

3. The tested random sample was quite representative. We were luckier than we deserved!

4. The Title I cases for whom data were available reinforced many observed characteristics of children in the stipulated socio-economic strata. However, I think it only fair to note that there are few target schools as such in New Hampshire outside a few of the large communities. Many of these failed to test in any case. Thus Title I help may have been extended to needful children not necessarily from economically deprived homes. This is good in a state where the legislature apparently cares so little for the welfare of our children.

Among such groups (as delineated by the available Title I data):

a. Boys fall behind girls most of the time in many measurable ways.

b. Learning ability of the Title I samples is lower than average by significant amounts and this is reflected in school performance. Grade 6 sample is better than Grade 4 and generally Grade 6 end-of-year performance is relatively better than Grade 4 at the same time period.

   It profits us nothing to argue about inherited vs environmentally derived cognitive skills. These kids need special recognition, special instructional materials, more individual attention including outside supportive services by way of mental health and reading clinics, more love and affection and thus less sense of failure and self-deprecation.

   Obviously these are personal opinions, but they are borne out by experience and common sense and are consistent with these data.

c. Relatively few of the Title I cases show clear signs of having correctible defects in learning skills compared to being just a slow moving group. Our data were not analyzed to maximize the chances of discovering such disabilities since, in this initial report all tested Title I cases are combined.

   More effort should be expended to search out these children with special cognitive learning blockings by adequate diagnoses and to provide the kind of corrective instruction that might be called the "prescription education" to emphasize its relevancy to individual needs.

The U.S. Office of Education should plan more carefully to stimulate these two kinds of efforts, i.e., better instruction for slow learners with reasonable goals; and diagnostic identification and remediation for the educationally handicapped. The nonsensical specification that all be brought up to grade, whatever that means, should be buried enough so it shall not confuse the issue any longer.

Packaged, debugged programs, thoroughly field tested, should be recommended by the USOE with especial care for the pupil accounting aspects of the program. Provision should be made for cumulative data files. No danger of the pygmalion phenomenon (in reverse) need worry us if teachers will learn to see children as individuals in a competitive society that had just better realize that it takes all kinds of people to make the world go around. The self-fulfilling prophecy no-

tion is an insult to the intelligence as well as the good intentions of the teaching staff. Given half an opportunity, there are few time-tested teachers (those staying in the profession because they love children) who will not welcome new ways, new aids, and new horizons to enable them to help all children learn.

In all this "evaluation" effort the child is almost forgotten. Has he been informed as to why we test - and retest? Has he even seen the results of his efforts in situ in terms of questions answered or not answered correctly? Has the notion of testing as one way he can get across to his teacher just what he does and does not know ever been thought of by the teacher to say nothing of having been communicated to him? If so, little evidence of this has reached this level of activity.

Holt's philosophy in "Why Children Fail" is largely beside the point. "Crisis in the Classroom" comes much closer to the truth.

Not enough has been said about evaluating performance project by project. The problem in New Hampshire is the small size of the local administrative units and thus the small N's one has to deal with.

Comparison of distributions in terms of central tendency statistics or percentiles is not satisfactory. The best plan would appear, at this moment in time, to be a pupil-by-pupil evaluation where pupils would be studied agsinst before-after bivariates of the most relevant tests. The "most relevant test" in reading might be a standardized battery, but might constitute selected material from such a battery or two forms of the reading test given within the same week and repeated at the end of the period of instruction.

The greatest chance to show dramatic changes is to work in an area such as math where individual item changes might be observed. The curriculum valid material selected would be that suitable for the Title I group and not just the test ordinarily used at a particular grade. To use the nomenclature commonly used in this report, the spring test answer document might be the scoring key for the fall document. The scoring might involve a multi-step procedure to determine:

a. the number and character of identical items answered consistently.

b. the number going from wrong to right and vice versa.

c. the percent of the group answering selected items considered needful of mastery as an hierarchical step-up toward an eventual goal of mastery of essential skills.

In math at any grade or developmental level one must start with demonstrated skill in number manipulation (computation). This must precede much attention to problem solving since competency in computation sets a go-no go limit on applications involving the specific skills. A careful study of the item analysis information from the Metropolitan Math tests recently released will reveal woeful lack of skill in number manipulation at any developmental level. Any nonsense about computation skills not being necessary in this day and age because of the advent of computers is sheerest irrelevancy. Problem solving will always be necessary and number manipulation is its prerequisite.

Not enough attention has been paid to discovery of the item types that are most suitable in before-after comparisons. The writer could fill another report on this subject, but will content himself with one generalization - to wit, item types relatively much freer of guessing than present item forms are absolutely essential and are in our grasp if the local groups demand such tests.

These are merely avenues for exploration, but at least they are not the fruitless blind pecking activity that best describes the continued reliance on total score comparisons, especially when fallacious methods of interpreting (e.g. grade equivalents) are applied to the data.

My last word

Title I programs should not be bandaids on a bad bruise, but preventive education that never lets the bruising situation occur because insightful, responsive, and ingenious people are trying hard with some special help to meet the needs of each child. Perhaps there should be a Title II program for the highly advantaged who, in most school situations, are equally put upon!

Appendix A

New Hampshire Statewide Testing Program
Intercorrelations of Stanford Achievement Test and
Otis-Lennon Mental Ability Test
Fall - 1968

Grade 4
SAT: Intermediate I, Form X
OLMAT:Elementary II, Form J

Test Name and Number		1	2	3	4	5	6	7	8	9	10	11	12	13
Word Meaning	1	1.00												
Para. Meaning	2	.75	1.00											
Spelling	3	.66	.68	1.00										
Word Study Sk.	4	.64	.62	.65	1.00									
Language	5	.70	.71	.69	.72	1.00								
Arith. Comp.	6	.35	.42	.40	.41	.47	1.00							
Arith. Conc.	7	.55	.59	.49	.57	.62	.51	1.00						
Arith. Appl.	8	.57	.61	.50	.58	.62	.49	.71	1.00					
Social Stud.	9	.68	.70	.56	.61	.67	.38	.64	.66	1.00				
Science	10	.74	.74	.62	.64	.70	.37	.60	.64	.75	1.00			
Comp. Prog.	11	.82	.87	.71	.73	.80	.58	.78	.80	.83	.81	1.00		
O-L MAT	12	.76	.76	.66	.70	.75	.44	.68	.69	.74	.76	.91	1.00	
I.Q.	13	.72	.71	.63	.68	.73	.42	.64	.65	.68	.71	.89	.94	1.00

Grade 6
SAT:Intermediate II, Form X
OLMAT:Elementary II, Form K

Test Name and Number		1	2	3	4	5	6	7	8	9	11	12	13
Word Meaning	1	1.00											
Para. Meaning	2	.80	1.00										
Spelling	3	.64	.68	1.00									
Language	4	.71	.76	.69	1.00								
Arith. Comp.	5	.42	.50	.48	.55	1.00							
Arith. Conc.	6	.60	.63	.50	.66	.60	1.00						
Arith. Appl.	7	.62	.67	.53	.68	.60	.77	1.00					
Social Stud.	8	.73	.77	.55	.73	.49	.70	.74	1.00				
Science	9	.74	.78	.53	.70	.43	.62	.67	.79	1.00			
Comp. Prog.	11	.84	.89	.68	.83	.66	.81	.84	.86	.80	1.00		
O-L MAT	12	.74	.76	.61	.78	.51	.68	.72	.75	.73	.90	1.00	
I.Q.	13	.71	.74	.60	.76	.51	.68	.68	.72	.68	.90	.94	1.00

Appendix B

TABLE 29
Correlations Between Otis-Lennon and Stanford Achievement Test

Grade	N	Otis-Lennon MAT Level	Otis-Lennon MAT Raw Score DIQ Mean	Otis-Lennon MAT Raw Score DIQ S.D.	Stanford Achievement Test: 1964 Edition* Level	Subtest	Raw Score Mean	Raw Score S.D.	
1	407	Prim. II	43.13	6.74	Prim. I	Word Reading	21.33	6.48	.52
			103.64	14.23		Paragraph Meaning	20.81	8.92	.47
						Vocabulary	21.79	6.31	.62
						Spelling	11.41	5.46	.42
						Word Study Skills	36.58	8.37	.54
						Arithmetic	38.09	11.93	.57
3	580	Elem. I	54.93	11.00	Prim. II	Word Meaning	24.89	5.42	.62
			103.78	13.57		Paragraph Meaning	47.28	9.54	.60
						Science-Social Studies	23.81	5.35	.56
						Spelling	20.58	6.52	.44
						Word Study Skills	48.23	11.25	.57
						Language	47.53	9.81	.59
						Arith. Computation	37.18	9.53	.50
						Arith. Concepts	29.83	8.37	.57
5	619	Elem. II	51.56	14.06	Inter. II	Word Meaning	24.26	8.52	.7.
			103.85	14.77		Paragraph Meaning	33.35	11.30	.78
						Spelling	28.15	9.48	.62
						Language	85.13	18.12	.78
						Arith. Computation	16.07	6.20	.60
						Arith. Concepts	14.38	5.57	.73
						Arith. Applications	17.94	6.89	.75
						Social Studies	39.43	12.26	.74
						Science	31.56	9.71	.75
7	607	Inter.	45.46	14.80	Adv.	Paragraph Meaning	32.01	11.85	.80
			105.34	14.20		Spelling	28.51	12.05	.63
						Language	94.15	17.05	.80
						Arith. Computation	19.06	7.79	.67
						Arith. Concepts	18.27	7.25	.74
						Arith. Applications	14.02	4.86	.67
						Social Studies	46.39	12.52	.80
						Science	33.91	9.33	.70
9	88	Inter.	53.75	15.17	H. S.	English	47.31	16.29	.83
			103.52	13.24		Numerical Competence	27.03	7.97	.79
						Mathematics	19.47	6.63	.78
						Reading	31.31	11.10	.83
						Science	32.05	9.44	.74
						Social Studies	28.88	7.77	.72
						Spelling	27.92	9.95	.82
11	84	Adv.	45.86	13.81	H. S.	English	53.73	13.95	.76
			101.92	12.27		Numerical Competence	30.12	9.09	.79
						Mathematics	23.60	8.74	.79
						Reading	37.25	9.88	.82
						Science	35.05	8.82	.68
						Social Studies	33.44	9.92	.74
						Spelling	32.89	11.40	.53

*Stanford administered approximately 2 months after Otis-Lennon.

From the Otis-Lennon Mental Ability Test Technical Handbook
Reproduced by permission of the publisher